(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 May 2003 (22.05.2003)

PCT

(10) International Publication Number
WO 03/042774 A2

(51) International Patent Classification⁷: G06F

(21) International Application Number: PCT/IB02/05827

(22) International Filing Date: 13 November 2002 (13.11.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/338,578 13 November 2001 (13.11.2001) US

(71) Applicant: CAPRION PHARMACEUTICALS INC. [CA/CA]; 7150 Alexander Fleming, Montreal, Quebec H4S 2C8 (CA).

(72) Inventors: **AGUILERA, Frank, Reinaldo, Morales**; 2832 Lippe, Montreal, Quebec H4R 1M1 (CA). **FAUBERT, Denis**; 10563 Georges-Baril Street, Montréal, Quebec H2C 2N4 (CA). **BOULOS, Marguerite**; 7131b de Normanville, Montreal, Quebec H2S 2C4 (CA). **TSANG, John, Shing-Chun**; 3875 Rue Broadway #6, Lachine, Quebec H8T 1T5 (CA). **HU, Michael**; 4998 De Maisonneuve O., Apt. 611, Montreal, Quebec H3Z 1N2 (CA). **OSTERMANN, Joachim, Bernhard**; 48 Tunstall Avenue, Senneville, Quebec H9X 1T2 (CA). **KEARNEY, Paul, Edward**; 41 Bruce Avenue, Montreal, Quebec H4Z 2E1 (US). **THIBAULT, Pierre**; 218 des Explorateurs, Aylmer, Quebec J9J 1M9 (CA).

(74) Agents: **ARMSTRONG, R., Craig** et al.; Borden Ladner Gervais LLP, World Exchange Plaza, 100 Queen Street, Suite 1100, Ottawa, Ontario K1P 1J9 (CA).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: MASS INTENSITY PROFILING SYSTEM AND USES THEREOF

(57) Abstract: The present invention is directed to computer automated methods and systems for identifying and characterizing biomolecules in a biological sample. Mass spectrometry measurements are obtained on biomolecules in a sample. These measurements are analyzed to determine the abundance of the biomolecules in the sample, and the abundance measurements are couple with one or more distinguishing characteristics of biomolecules they are associated with, thereby permitting computer-mediated comparison of abundances of biomolecules from multiple biological samples.

MASS INTENSITY PROFILING SYSTEM AND USES THEREOF

BACKGROUND OF THE INVENTION

The invention relates to the fields of mass spectrometry, bioinformatics, and computational molecular biology. In particular, this invention relates to the automation of biomolecule quantification.

Genomic and proteomic research efforts in recent years have vastly improved our understanding of the molecular basis of life at a global cellular and tissue scale. In particular, it is increasingly clear that the temporal and spatial expression of an organism's biomolecules is responsible for life's processes -- processes occurring in both health and in sickness. Science has progressed from understanding how genetic defects cause hereditary disorders, to an understanding of the importance of the interaction of multiple genetic defects together with environmental factors in the etiology of complex medical disorders, such as cancer. In the case of cancer, scientific evidence demonstrates the key causative roles of altered expression of, and multiple defects in, several pivotal genes and their protein products. Other complex diseases have similar molecular underpinnings. Accordingly, the more complete and reliable a correlation that can be established between expression of an organism's biomolecules and healthy or diseased states, the better diseases can be diagnosed and treated. Methods that permit efficient and rapid quantification of biomolecule expression from biological samples that may contain tens of thousands of different biomolecules of a particular type (e.g., protein, lipids, nucleic acids, carbohydrates, metabolites, and combinations thereof) are necessary to provide the best possible chance to determine such correlations. For example, proteomic data reflects the true expression levels of functional molecules and their post-translational modifications, which cannot be accurately predicted from other data types such as gene expression profiling.

A central goal of proteomics, which involves the systematic identification and characterization of proteins in a sample, is to be able to compare the protein composition between two or more samples. Critical to achieving this goal is the ability to identify all the proteins that are present in only one sample or type of sample and any proteins that are present in several samples or types of samples but differ in abundance. The complexity and dynamic nature of the proteomes of living beings, however, as well as limitations in sample quantity and stability, provide enormous challenges in identifying the amino acid sequence and the source protein(s) of the peptidic material present in a sample, in quantifying the relative abundance of the different peptides or source proteins present in the sample, and in providing complete enough data about a sample to produce an accurate snapshot of the proteome. The complexity of the proteomes and of these tasks further makes performing such proteome-wide analyses and comparisons difficult to accomplish in a reasonable time frame.

Comparability itself is also an issue. The methods by which two biomolecules are judged to be the same or comparable depend on the methods employed to identify a particular biomolecule within a field of biomolecules and the completeness of the data gathered from a sample. Usually, a comparison of all the proteins in a sample is accomplished by two-dimensional (2D) gel electrophoresis, which resolves a complex protein mixture into hundreds or thousands of spots, which have characteristic migratory positions for particular proteins. Gel patterns can be directly comparable with correction of migration variables if the gel and sample were properly prepared and run, but gel reproducibility is quite variable from lab to lab or even with different lots of ampholines. Each spot, in theory, represents one protein, and the intensity of each spot is taken as a measure for the amount of the protein present. The protein that is present in this spot can then be more fully identified by mass spectrometry or other methods; however, the further identification of a single protein spot, let alone the whole field of spots, can involve considerable time, effort, and expense. The 2D electrophoresis approach also has several other drawbacks, the most important of which is the difficulty of identifying membrane proteins. In general, 2-D electrophoresis has problems with the exclusion of highly hydrophobic molecules, and with the detection of highly charged (very acidic or very basic) molecules, as well as of very small or very large molecules. In addition, the detection of low or even moderate abundance proteins is difficult and may require that several gels be run to collect enough material for sequence analysis. 2D gel spots can also be quite large, which dilutes the protein over a large part of the gel, rendering detection and accurate quantification of proteins more difficult. Additionally, co-migration of proteins, particularly of closely related or variant proteins, can interfere with both proper identification and quantification of the specific proteins.

One-dimensional (1D) gel electrophoresis, on the other hand, is a generally applicable tool to separate proteins that at least allows the study of both soluble and membrane proteins. However, when complex mixtures of proteins are analyzed, only 50 to 100 protein bands are typically detectably produced in the separation, and a single band in a 1D gel may, therefore, contain more than a single protein. For this reason, the intensity of one band does not typically reflect the abundance of a single protein in the sample, and identification likewise becomes more problematic. Mass spectrometry, for example, of a single band will lead to the identification of not just one but several (e.g. 10 to 20) proteins that are present in the band at different concentrations.

Mass spectrometry itself is another method of choice for analyzing complex mixtures of molecules, such as the contents of cells, or cellular components. When combined with appropriate methods of chromatography to allow separation and purification of biomolecules, mass spectrometry provides a start point for producing and analyzing data for the identification and quantification of biomolecules, and for patterns that liken or distinguish different samples. At its most basic, mass spectrometry produces data about the mass of biomolecules, and their intensity (ion counts) for a particular scan. Fragmentation patterns for specific molecules can also be produced, but these characteristic spectra, which can be used to further identify the molecule, are unlinked to the quantitative data (ion counts) produced in the initial scan. Secondary efforts are required to derive structural information from this basic data, or, in the case of polymers such as DNA or proteins, to obtain sequence information from the fragmentation patterns, to determine the source protein from the sequence information, and to couple sequence/identity information to quantification data.

One quantitative mass spectrometric technique relies on coupling different isotopic tags to the peptides of each sample to be analyzed. An example of this methodology is referred to as isotope-coded affinity tag (ICAT) (see Han et al. (2001) Nat. Biotechnol. 19: 946 - 51 (PMID: 11581660)). This method consists of derivatizing proteins, such as with alkylating agents containing a reactive group specific to cysteine residues, a linker chain, and a biotinylated moiety. The alkylating agent includes a light and a heavy version corresponding to 8 hydrogen atoms (light) or 8 deuterium atoms (heavy) in the linker chain. When comparing two samples, all the peptides from one sample are tagged with the light tag, and all the peptides from the other sample are tagged with the heavy tag. Both samples are then mixed, digested with trypsin, and analyzed simultaneously. In the mass spectrum, ions pairs that correspond to the same peptide but differ by the exact mass difference (8 Da) between the heavy and light tag are then identified. These ions then correspond to the same peptide but are derived from the two different samples. This method allows for the direct comparison of the abundance of corresponding peptides from the two samples. Despite permitting direct comparison of samples, this technique generally has the limitation that all peptides containing cysteine residues must be chemically modified before they are analyzed. Such modifications come at an additional expense in both money and time. They can also have a cost in accuracy if the reaction does not go to completion, or the delays due to processing time lead to protein degradation. Furthermore, the chemical modification requires the presence of a specific amino acid, cysteine, in the peptide, which means that the majority of peptides are not suitable for the analysis. This requirement greatly reduces the applicability of this approach to a wide range of proteins. The ICAT approach can also generate interfering intensities from biotinylated fragment ions in MS/MS experiments, hampering the ability to determine peptide sequence information.

Another labeling method uses light and heavy isotopes of water. Tryptic peptides from different protein pools are labeled at the C-terminus with ^{16}O and ^{18}O water. This method has been used to distinguish between b- and y-type fragment ions in MS/MS experiments (see Schevshenko et al. (1997) Rapid Commun. Mass Spectrom. 11: 1015 - 1024). The method has also been used for monitoring the differential expression of proteins in two serotypes of adenovirus (see Yao et al. (2001) Anal. Chem. 73: 2836 - 2842). As above, protein pools are digested separately, labeled, and combined for analysis by mass spectrometry. Expression profiles are then obtained based on the ratio of heavy to light ions. This method also requires that the peptides or proteins be labeled before analysis, and thus, like ICAT may suffer from incomplete reactions, substrate insusceptibility, extra cost, and extra preparation time made all the more costly by the possible detriment to limited and potentially unstable samples. These issues are exacerbated by the additional challenges of preparing such samples from living organisms.

Existing methods for biomolecule identification or characterization are therefore in need of improvement in their ability to perform rapid, accurate, automated, and economical as well as qualitative, quantitative, and specific determinations of the components of a biological sample. For example, there exists a need for methods using mass spectrometry to determine the abundance of peptides in a sample that does not require chemical modification of the peptides. Furthermore, there is a continuing and significant need to be able to readily compare the relative abundances of proteins between biological samples, and to identify and characterize proteins as targets for drug discovery. The present invention fulfills these needs and further provides other related advantages.

SUMMARY OF THE INVENTION

The present invention features computer automated methods and systems for identifying and characterizing biomolecules in a biological sample. In these methods, mass spectrometry measurements are obtained on biomolecules in a sample. These measurements are then analyzed by the methods described herein to determine the abundance of the biomolecules in the sample, and the abundance measurements are coupled with one or more distinguishing characteristics of biomolecules they are associated with, thereby permitting computer-mediated comparison of abundances of biomolecules from multiple biological samples. We refer to this technology as "MIPS" or mass intensity profiling system. This automated MIPS technology for screening biological samples and comparing their mass intensity profiles permits rapid and efficient identification of individual biomolecules whose presence, absence, or altered expression is associated with a disease or a condition of interest. Such biomolecules (for example, proteins) are potentially useful as therapeutic agents, as targets for therapeutic intervention, or as markers for diagnosis, prognosis, and evaluating response to treatment. MIPS technology also permits rapid identification of sets of biomolecules whose pattern of expression is associated with a disease or condition of interest; such sets of biomolecules provide a collection of biological markers for potential use in diagnosis, prognosis, and evaluating response to treatment.

In one aspect, the invention features a method for determining an abundance of a biomolecule in a biological sample. In general, the method includes the steps of providing a biological sample containing a plurality of biomolecules; generating a plurality of ions of the biomolecules; performing mass spectrometry measurements on the plurality of ions, thereby obtaining ion counts for the biomolecules; assigning an ion to a biomolecule; and integrating the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

In particular, the invention features methods and systems for the determination of the abundance of peptides in a sample, but the following methods may be applied to other biomolecules as well. The invention further features methods and systems that can be used to compare expression levels of peptides between two or more samples, as well as methods to compare expression levels of proteins between two or more samples. These methods are based on the analysis of data from mass spectrometry.

In various embodiments, MIPS can be used to determine the abundance of peptidic material present in one or more LC/MS scans. MIPS may be used to determine the abundance for all the precursors present, all precursors which have been matched with spectra generated by MS/MS, or precursors as limited by a list (inclusion or exclusion), a query, or set of queries.

In one embodiment, MIPS can be used to calculate the abundance of one or more peptides without a determination of amino acid sequence or full-length protein identification. Alternatively, MIPS can be used to calculate the abundance of one or more peptides with a determination of amino acid sequence but without full-length protein identification. In another embodiment, MIPS can be used to calculate the abundance of one or more peptides and perform peptide and full-length protein identification without prior amino acid sequence determination, through methods including, but not limited to, peptide mass fingerprinting (see, for example, Cottrell (1994) Pept. Res. 7: 115 - 24). In still another embodiment, MIPS can be used to calculate the abundance of one or more peptides, to determine the amino acid sequence and to identify the full-length protein of which a peptide is a constituent. The above methods may also be modified to include determining an amino acid sequence but not relying upon this information in determining the identity of a full-length protein. The abundance of a full-length protein so identified can be calculated, based on the abundance of one or more of its constituent peptides. These methods can also be used to match or compare the abundance of the same peptide or protein between two or more samples or the abundance of different peptides or proteins in the same or a different sample. Comparisons between samples may include comparing data from a patient to a reference, manipulated, representative, combined, or theoretical samples.

In various embodiments, MIPS can be used to identify one or more constituent peptides of an identified full-length protein, and to calculate the abundance of one or more of the identified constituent peptides. Such abundance data may be used in turn to calculate or recalculate the abundance of a full-length protein.

In another embodiment, MIPS can be used to query the abundance of one or more peptides or full-length proteins in one or more samples, with or without prior calculation of said abundances, and with or without prior identification of the one or more peptides or proteins. Queries may be manually entered or part of an automated process, such as a list of ions to exclude from MS/MS. Queries may take place prior to data acquisition and wholly or in part determine the data collected, or they may occur post-data acquisition to mine data.

In various embodiments, the calculation of full-length protein abundance may be based upon the abundance of one or more constituent peptides. Such peptides may have been previously identified, or they may be sought (automatically or manually) and their abundances calculated based on the predicted peptides for the particular full-length protein.

In various embodiments, the calculation of peptide abundance may be absolute or relative. In general, abundance is determined by a sum of ion counts based on a consistent choice within a sample, for example, a subset of charge states, isotopes, modified states, or a combination thereof.

In various embodiments, comparisons may be to the same or a different peptide or protein. Comparisons may also be to homologous or analogous peptides or proteins, individual peptides or proteins, or peptides or proteins representative of a family or group. Comparisons may also be to unrelated peptides or proteins. Data may be extrapolated and statistically analyzed for individual peptides and/or proteins, and/or groupings thereof. Data used for comparison may be from within the same set of sample data, and/or from one or more other sets of data including, but not limited to, reference, manipulated, representative, combined, and/or theoretical samples.

In various embodiments, matching a peptide or peptides, protein or proteins can be based on experimental or theoretical retention time. Theoretical retention time may be calculated to take into account error, charge state, theoretical and/or adjusted charge state, or isotopes, and they may also be adjusted using one or more internal standards, including those exogenous to a sample. Other methods include, but are not limited to, pattern monitoring, such as isotope pattern monitoring (peptide recognition is based on the isotopic model of a given peptide (height, width and distance of peaks)). Such matching may be for purposes including, but not limited to, querying data, comparisons, or uniting data from separate data sources including, but not limited to, separate scans, spectra, sequence information, annotations, and combinations thereof.

In various embodiments, a peptide or protein in a sample for which an abundance can be calculated may be used to generate a list of one or more peptides or proteins, which may in turn be combined with other lists or used directly or indirectly for querying, matching, or governing data gathering, such as selection for spectra determination in further analysis of the same or another sample.

The invention further features a computer implemented method for determining abundance of a biomolecule in a biological sample. The computer implemented method generally includes the steps of inputting mass spectrometry data comprising ion counts for a plurality of biomolecules; assigning an ion to a biomolecule; and integrating the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

In another aspect, the invention features a computer-readable memory that includes a program for determining abundance of a biomolecule in a biological sample including computer code that receives as input mass spectrometry data comprising ion counts for a plurality of biomolecules; computer code that assigns an ion to a biomolecule; and computer code that integrates the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

In yet another aspect, the invention features a system for determining abundance of a biomolecule in a biological sample including a mass spectrometry data input module that receives data including ion counts for a plurality of biomolecules; an ion assigning module responsive to the data input module, wherein the ion assigning module assigns an ion to a biomolecule; and an ion integrating module responsive to the ion assigning module, wherein the ion integrating module integrates ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample. In one embodiment, the system includes a processor and a memory coupled to the processor, wherein the memory encodes the data input module, the ion assigning module, and the ion integrating module.

In another aspect, the invention features a method for displaying information on abundance of a biomolecule in a biological sample to a user including the steps of inputting mass spectrometry data comprising ion counts for a plurality of biomolecules; assigning an ion to a biomolecule; integrating the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample; and displaying the abundance of the biomolecule. In one embodiment, the method further includes storing the abundance of the biomolecule in a memory.

In various embodiments of any of the aforementioned aspects, the biomolecule is underivatized or unlabeled. The biomolecule may also be cleaved biomolecule. In preferred embodiments, the biomolecule is cleaved with an enzyme. In general, however, the methods do not require modification other than cleavage, such as isotope-labeling or akylation, of the biomolecules, i.e., cleaved biomolecules may be underivatized or unlabeled. In yet other preferred embodiments, the invention further features assaying two or more biological samples. The invention, if desired, features the inclusion of one or more internal standards in the biological sample. In still another embodiment, a computer procedure assigns the ion to the biomolecule by calculating an uncharged mass for the ion. Alternatively, ions may be assigned to biomolecules through mass fingerprinting, e.g., peptide mass fingerprinting. In yet another embodiment, a computer procedure integrates ion counts of the ions corresponding to the biomolecule. Preferably, the integration is over one or more charge states, isotopes, scans, fragments of the biomolecule, fractions of a separation, or a combination thereof. In other embodiments, the invention further features separating the plurality of biomolecules prior to MS analysis. Typically, such separation is carried out using standard methods known in the art. These methods include, without limitation, chromatography, electrophoresis, immunoisolation (e.g., using magnetic beads), or centrifugation. The retention time of an ion may be corrected using one or more internal standards.

In various other embodiments of any of the aforementioned aspects, the biomolecule is typically a protein or modified protein. Preferably, the protein is obtained from an isolated organelle. Exemplary isolated organelles include, without limitation, mitochondria, chloroplasts, ER, Golgi, endosomes, lysosomes, phagosomes, peroxisomes, secretory vesicles, transport vesicles, nuclei, and plasma membrane. Proteins obtained from other cellular components are also useful in the invention. These proteins include cytosolic or cytoskeletal proteins.

In preferred embodiments, mass spectrometry measurements are obtained to gather structural or sequence information of an ion of the biomolecule, e.g., through MS/MS analysis. Biomolecules or ions thereof may be selected for structural or sequence analysis (e.g., MS/MS analysis) by a query. In one embodiment, an inclusion or exclusion list is used to determine which ions will be subjected to structural or sequence analysis. The methods and systems of the invention further feature the use of a computer procedure to identify a protein comprising the sequence of the ion from a database. Exemplary procedures include Mascot®, Protein Lynx Global Server, SEQUEST®/TurboSEQUEST, PEPSEQ, SpectrumMill, or Sonar MS/MS. Exemplary databases that are searched using such procedures include the Genbank®, EMBL, NCBI, MSDB, SWISS-PROT®, TrEMBL, dbEST, or Human Genome Sequence database. Moreover, the methods and systems include a computer procedure that assigns the ion to the protein identified from a database.

In various other embodiments of any of the aforementioned aspects, the invention features calculating an abundance of the biomolecule relative to a control biological sample and calculating abundances of a plurality of the biomolecules relative to a control biological sample. Typically, abundance measurements of a set of biomolecules are used to diagnose a disease or condition. Additionally, abundance is used to determine a biomolecule to target with a drug. Such targets are identified by evaluating an increase or decrease in abundance or the presence or absence of a biomolecule in the biological sample relative to a control sample. Abundance of a biomolecule may also be used to determine an amount of an isoform of a biomolecule, or of a naturally occurring modification of a biomolecule.

By "biomolecule" is meant any organic molecule that is present in a biological sample, including peptides, polypeptides, proteins, post-translationally modified peptides or proteins (e.g., glycosylated, phosphorylated, or acylated peptides), oligosaccharides, polysaccharides, lipids, nucleic acids, and metabolites.

By "biological sample" (or "sample") is meant any solid or fluid sample obtained from, excreted by, or secreted by any living organism, including single-celled micro-organisms (such as bacteria and yeasts) and multicellular organisms (such as plants and animals, for instance a vertebrate or a mammal, and in particular a healthy or apparently healthy human subject or a human patient affected by a condition or disease to be diagnosed or investigated). A biological sample may be a biological fluid obtained from any location (such as blood, plasma, serum, urine, bile, cerebrospinal fluid, aqueous or vitreous humor, or any bodily secretion), an exudate (such as fluid obtained from an abscess or any other site of infection or inflammation), or fluid obtained from a joint (such as a normal joint or a joint affected by disease such as rheumatoid arthritis). Alternatively, a biological sample can be obtained from any organ or tissue (including a biopsy or autopsy specimen) or may comprise cells (whether primary cells or cultured cells) or medium conditioned by any cell, tissue or organ. If desired, the biological sample is subjected to preliminary processing, including preliminary separation techniques. For example, cells or tissues can be extracted and subjected to subcellular fractionation for separate analysis of biomolecules in distinct subcellular fractions, e.g., proteins or drugs found in different parts of the cell. A sample may be analyzed as subsets of the sample, e.g., bands from a gel.

By "assigning an ion to a biomolecule" is meant specifying a biomolecule from which an ion observed in a mass spectrum was generated. The ion may be assigned to a biomolecule or a fragment thereof. Such assignments may be based, for example, on the molecular mass, or other physicochemical characteristic. The assignment can also be made on the basis of determining the molecular mass of the ion and matching that mass with a known biomolecule or on the basis of data, e.g., from MS/MS, that identifies structural or sequence information about the ion, which may be used to search a database.

By "uncharged mass" is meant the mass of the neutral charge state of the biomolecule or a fragment thereof from which an ion is generated.

By "integrating the ion counts of a biomolecule" is meant summing ion counts for data within a defined range of m/z values. The phrase also refers to summing integrated ion counts of two or more ions. For example, ions that are found in different charge states, isotopes, fractions of a separation, scans, or fragments of a biomolecule may be integrated.

By the term 'protein" is meant any polymer of two or more individual amino acids linked via a peptide bond that forms when the carboxyl carbon atom of the carboxylic acid group bonded to the alpha-carbon of one amino acid (or amino acid residue) becomes covalently bound to the amino nitrogen atom of amino group bonded to the alpha-carbon of an adjacent amino acid. The term "protein" is understood to include the terms "polypeptide" and "peptide" (which, at times, may be used interchangeably herein) within its meaning, as well as post-translational modifications and fragments thereof. In addition, proteins comprising multiple polypeptide subunits (e.g., insulin receptor, cytochrome b/c1 complex, and ribosomes) or other components (for example, an RNA molecule) will also be understood to be included within the meaning of "protein" as used herein. Similarly, fragments of proteins and polypeptides are also within the scope of the invention and may be referred to herein as "proteins," "polypeptides," or "peptides," "tryptic peptides", or "cleavage fragments." "Constituent peptides" are peptides whose sequence is a linear subset of the sequence of a larger peptide or full-length protein. As a group, the "constituent peptides" for a particular protein would be a set or subset of those that make up the protein. Usually, this is a subset limited to particular cleavage fragments, such as the set of tryptic peptides that make up a protein. A "full-length protein" refers to a protein encoded by and translated from a messenger RNA (mRNA), and post-translational modifications thereof. Full-length proteins may be identified through database searching via computer procedures as described herein.

By "precursor" is meant a biomolecule, e.g., a potential peptide or protein or one of unknown sequence or identity. Generally it refers to potential peptides in mass spectrometry survey scan data prior to secondary identification efforts, such as sequencing by MS/MS. "Precursors" are frequently identified by comparing their masses or their retention times. Such retention times may be experimental or theoretical. Theoretical retention times are frequently corrected, where one or more internal standards are used to make retention times comparable between samples. Predicted retention times may be used to seek precursors within a scan. "Precursor" is frequently used interchangeably with "peptide," and it may be used to distinguish individual constituent peptides from full-length proteins.

By "scan" is meant a mass spectrum from a single sample. Each fraction of a separation that is measured results in a scan. If a biomolecule is located in more than one fraction analyzed, then the mass spectrum for the biomolecule is present in more than one scan.

By "fraction" is meant a portion of a separation. A fraction may correspond to a volume of liquid obtained during a defined time interval, for example, as in LC (liquid chromatography). A fraction may also correspond to a spatial location in a separation such as a band in a separation of a biomolecule facilitated by gel electrophoresis.

By "query" is meant a selection of a particular action. In one example of a query, ions may be subjected to MS/MS based on a list that is stored with the software. Alternatively, one can manually select ions to be subjected to MS/MS. This manual selection is also a query.

By an "underivatized" biomolecule or fragment thereof is meant a biomolecule or fragment thereof that has not been chemically altered from its natural state. Derivitization may occur during non-natural synthesis or during later handling or processing of a biomolecule or fragment thereof.

By an "unlabeled" biomolecule or fragment thereof is meant a biomolecule or fragment thereof that has not been derivatized with an exogenous label (e.g., an isotopic label or radiolabel) that causes the biomolecule or fragment thereof to have different physicochemical properties to naturally synthesized biomolecules

The methods and systems of the invention provide a number of significant advantages. For example, the methods and systems combine mass spectrometry and data analysis in a way that allows both the identification of proteins and a measure of their abundance. The methods thus allow a correlation of ion intensities of a biomolecule between sample sets in order to compare relative abundance of the original biomolecule from which the biomolecules are derived. Further, the methods are fully automated allowing a comprehensive comparison of an unlimited number of proteins from different data sets. This automation thus allows for protein abundances in two or more samples to be compared. The information from the entire mass spectrum can also be used to determine expression levels. Typically, a large amount of information that is present in the mass spectra is discarded, and only a subset, such as intensities of specific ions, or the sequence of specific peptides, or a list of peptide masses is analyzed. Finally, the use of automation greatly reduces the time necessary for analysis.

Other features and advantages of the invention will be apparent from the following drawings and detailed description, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates an exemplary embodiment of a computer system of this invention.

Figure 2 shows a flowchart of a method of determining an abundance of a biomolecule in a sample.

Figure 3 shows a flow chart for a Peptide Abundance Module. Solid rectangles represent processing components of MIPS, dashed rectangles represent processing components that are not within MIPS, and entries without a rectangle are data files.

Figure 4A shows a mass spectrum scan corresponding to a first elution time in a scan window.

Figure 4B shows a mass spectrum scan corresponding to a second elution time in a scan window.

Figure 4C shows a mass spectrum scan corresponding to a third elution time in a scan window.

Figure 4D shows a mass spectrum scan corresponding to a fourth elution time in a scan window.

Figure 5 show a mass spectrum illustrating the presence of two charge states.

Figure 6 shows a mass spectrum illustrating the separation between isotopes of a +3 charge state.

Figure 7 shows a flow chart for a Peptide Hunter Module.

Figures 8A-8I show results for three different m/z values from the HSP90 protein. Figures 8A-8C show the reconstructed ion chromatogram, survey scan, and MS/MS of the 924.4 m/z ion. Figures 8D-8F show the reconstructed ion chromatogram, survey scan, and MS/MS of the 757.4 m/z ion. Figures 8G-8I show the reconstructed ion chromatogram, survey scan, and MS/MS of the 621.8.

Figure 9 shows a table of the abundances of here m/z values of the HSP90 protein in five samples. The abundances were summed across several 1-D gel bands.

Figures 10A-10I show results for three different m/z values from the mutant desmin protein. Figures 10A-10C show the reconstructed ion chromatogram, survey scan, and MS/MS of the 561.3m/z ion. Figures 10D-

10F show the reconstructed ion chromatogram, survey scan, and MS/MS of the 558.3 m/z ion. Figures 10G-10I show the reconstructed ion chromatogram, survey scan, and MS/MS of the 466.7.

Figure 11 shows a table of the abundances of here m/z values of the mutant desmin protein in five samples. The abundances were summed across several 1-D gel bands.

Figures 12A and 12B show differential expression of HSP90 and mutant desmin in U937 cells using MIPS

Figure 13 shows a flow chart for a Peptide Abundance Module processing matched samples (N/T), such as matched normal and tumor tissue, or treated/untreated cells. Elitox is an information extraction procedure which extracts RT from an original MS-MS. PAM(IS) is a procedure used to acquire internal standard information (m/z values, intensity, and retention times). PAM Wrapper is a procedure that integrates the information contained within the precursor list, Elitox, and PAM IS. RT correction uses data derived from any internal standards (PAM IS, RT and m/z values) as well as RT from original MS-MS (Elitox) to predict the retention times (RT) from MS data in an LC-MS survey and provides m/z values and retention time correction. QC checks are quality checks to assure the process is running correctly, e.g., whether it needs to be adjusted or aborted based on expected data characteristics.

Figure 14 shows an estimation of a PAM Processing Time for a specific sample.

Figure 15 shows a flow chart for a Peptide Hunter Module processing multiple samples (a reference sample and a treated sample in this example), comprising a Peptide Hunter Module, a Peptide Merger Module (PMM), a Band Merger Module (BMM), and a Differential Abundance Module (DAM).

Figure 16 shows an LC-MS survey scan with expansion around 700 m/z on which PHM was run.

Figure 17 shows the m/z values for peptides identified by PHM, as well as their charge and calculated abundance. Compare, for example, the abundance of the +1 peptide at 701.3402 to the +3 peptide at 699.9900 in the expansion of Figure 16 to the data of this figure.

Figure 18 shows a scatter plot of abundance data (log scale) for predicted full-length proteins (clustered peptides) comparing proteins from a normal sample to a matched tumor sample produced using PAM. The Na/K ATPase protein corresponds to a known plasma membrane (PM) marker that is not expected to change expression between normal cells and tumor cells, while CEA (carcinoembryonic antigen) is a protein known to be up-regulated in tumors.

Figure 19 shows a scatter plot of individual peptide abundances (log scale) from a normal sample and a matched produced using PAM.

DETAILED DESCRIPTION OF THE INVENTION

The invention features methods utilizing mass spectrometry and software to measure the abundance of a biomolecule, qualitatively or quantitatively, or both. In one application, the methods and systems of the invention are used to compare a large number of peptides present in two or more samples in order, e.g., to determine variations in relative expression levels or to identify peptides for which ratios of relative expression are above or below pre-set values. Statistical analysis of expression profiles are then used to identify peptide markers, e.g., for disease diagnostics and drug discovery.

Biological Samples

Using the methods of the invention, an expression profile of a biomolecule is monitored in a biological sample. Exemplary biomolecules useful in the methods of the invention include any organic molecule that is present in a biological sample, e.g., peptides, polypeptides, proteins, post-translationally modified peptides (e.g., glycosylated, phosphorylated, or acylated peptides), oligosaccharides and polysaccharides, lipids, nucleic acids, and metabolites. Virtually any biological sample is useful in the methods of the invention, including, without limitation, any solid or fluid sample obtained from, excreted by, or secreted by any living organism, including single-celled micro-organisms (such as bacteria and yeasts) and multicellular organisms (such as plants and animals, for instance a vertebrate or a mammal, and in particular a healthy or apparently healthy human subject or a human patient affected by a condition or disease to be diagnosed or investigated). A biological sample may be a biological fluid obtained from any location (such as blood, plasma, serum, urine, bile, cerebrospinal fluid, aqueous or vitreous humor, or any bodily secretion), an exudate (such as fluid obtained from an abscess or any other site of infection or inflammation), or fluid obtained from a joint (such as a normal joint or a joint affected by disease such as rheumatoid arthritis). Alternatively, a biological sample can be obtained from any organ or tissue (including a biopsy or autopsy specimen) or may comprise cells (whether primary cells or cultured cells) or medium conditioned by any cell, tissue, or organ. If desired, the biological sample is subjected to preliminary processing, including preliminary separation techniques. For example, cells or tissues can be extracted and subjected to subcellular fractionation for separate analysis of biomolecules in distinct subcellular fractions, e.g., proteins or drugs found in different parts of the cell. Such exemplary fractionation methods are described in De Duve ((1965) J. Theor. Biol. 6: 33 - 59).

When analyzing proteins, a biological sample, if desired, is purified to reduce the amount of any non-peptidic materials present. Moreover, if desired, protein-containing samples are cleaved to produce smaller peptides for analysis. Cleavage of the peptides is generally accomplished enzymatically, e.g., by digestion with trypsin, elastase, or chymotrypsin, or chemically, e.g., by cyanogen bromide. The cleavage at specific locations in a protein allows the prediction of the masses of the smaller peptides produced if the sequences of these peptides are known. All samples that are to be compared typically are treated in the same manner.

A reference sample, if desired, is also included when performing the methods described herein. This reference sample typically includes known amounts of biomolecules or may be derived from a known source, e.g., a non-diseased tissue. The reference sample may be synthesized from known biomolecules. Additionally, unknown samples may be compared to the reference sample to determine a relative abundance. Reference samples may also be combined with other samples to act as internal standards where appropriate.

Separation of Biomolecules

A wide variety of techniques for separating any of the aforementioned biomolecules are well known to those skilled in the art (see, for example, Laemmli Nature 1970, 227:680-685; Washburn et al., Nat. Biotechnol. 2001, 19:242-7; Schagger et al., Anal. Biochem. 1991, 199:223-31) and may be employed according to the present invention.

In one application, the methods of the invention are used to study complex mixtures of proteins. By way of example, mixtures of proteins may be separated on the basis of isoelectric point (e.g., by chromatofocusing or isoelectric focusing), of electrophoretic mobility (e.g., by non-denaturing electrophoresis or by electrophoresis in the presence of a denaturing agent such as urea or sodium dodecyl sulfate (SDS), with or without prior exposure to a reducing agent such as 2-mercaptoethanol or dithiothreitol), by chromatography, including LC, FPLC, and HPLC, on any suitable matrix (e.g., gel filtration chromatography, ion exchange chromatography, reverse phase chromatography, or affinity chromatography, for instance with an immobilized antibody or lectin or immunoglobins immobilized on magnetic beads), or by centrifugation (e.g., isopycnic centrifugation or velocity centrifugation).

In some cases, two different peptides may have the same mass within the resolution of a mass spectrometer, rendering determination of abundances for those two peptides difficult. Separating the peptides before analysis by mass spectrometry allows for the resolution of the abundances of two peptides with the same mass. Although many spectra for the fractions of the separation may then be obtained, these spectra typically have a reduced number of ion peaks from the peptides, which simplifies the analysis of a given spectrum.

In one embodiment, a mixture of proteins is separated by 1D gel electrophoresis according to methods known in the art. The lane containing the separated proteins is excised from the gel and divided into fractions. The proteins are then digested enzymatically. The peptides produced in each fraction are then analyzed by mass spectrometry. In another embodiment, peptides are separated by 2D gel electrophoresis according to methods known in the art. The proteins are then digested enzymatically, and the digested peptides produced in each fraction are then excised and analyzed by mass spectrometry. In still another embodiment peptides are separated by liquid chromatography (LC) by methods known in the art, including, but not limited to, multidimensional LC. LC fractions may be collected and analyzed or the effluent may be coupled directly into a mass spectrometer for real-time analysis. LC may also be used to separate further the fractions obtained by gel electrophoresis. Recording the retention time (RT) of a peptide in LC enables the identification of that peptide in multiple fractions. This identification is typically useful for obtaining an accurate abundance. In any of the above embodiments, a given peptide may be present in more than one fraction depending on how the fractions were obtained.

Mass Spectrometry

Exemplary methods for analyzing biomolecules using mass spectrometry techniques are well known in the art (see Godovac-Zimmermann et al. (2001) Mass Spectrom. Rev. 20: 1 - 57 (PMID: 10344271); Gygi et al. (2000) Proc. Natl. Acad. Sci. U.S.A. 97: 9390 - 9395 (PMID: 10920198)).

In applications involving peptides, the peptides are ionized, e.g., by electrospray ionization, before entering the mass spectrometer, and different types of mass spectra, if desired, are then obtained. The exact type of mass spectrometer is not critical to the methods disclosed herein. For example, in a survey scan, mass spectra of the charged peptides in a sample are recorded. Furthermore, the amino acid sequences of one or more peptides may be determined by a suitable mass spectrometry technique, such as matrix-assisted laser desorption/ionization combined with time-of-flight mass analysis (MALDI-TOF MS), electrospray ionization mass spectrometry (ESI MS), or tandem mass spectrometry (MS/MS). In a MS/MS scan, specific ions detected in the survey scan are selected to enter a collision chamber. The ability to define the ions for MS/MS allows data to be acquired for specific precursors, while potentially excluding other precursors. The ions may be defined by a predetermined list or by a query. Lists may be inclusion lists (i.e., ions on the list are subjected to MS/MS) or exclusion (i.e., ions on the list are not subjected to MS/MS). The series of fragments that is generated in the collision chamber is then analyzed again by mass spectrometry, and the resulting spectrum is recorded and may be used to identify the amino acid sequence of the particular peptide. This sequence, together with other information such as the peptide mass, may then be used, e.g., to identify a protein. The ions subjected to MS/MS cycles may be user defined or determined automatically by the spectrometer.

Mass Intensity Profiling System (MIPS)

Software to analyze mass spectra is typically used to identify the biomolecule from which an ion was derived. A mass spectrum, however, also includes information on the relative intensity of an ion as reflected by its corresponding ion intensity (e.g., ion counts per second). Moreover, a mass spectrum typically includes a large amount of information corresponding to ion intensities from several charge states or several isotopes of a biomolecule. As is described herein, an automated approach allows the processing of mass spectra recorded for one or more samples so that a comprehensive characterization of the biomolecules in that sample is achieved.

As is described in more detail below, the software used in the methods herein automatically identifies ion signals (e.g., peptide ion signals) in the spectra. A biomolecular abundance measure is then calculated by a variety of methods. For example, the intensity of a specific subset of biomolecular ions (such as all ions with +2 charge), the intensities of all ions derived from one isotope of a biomolecule (such as only the ^{12}C, ^{1}H, ^{14}N, and ^{16}O-containing biomolecules), the intensities of a certain subset of isotopes of a biomolecule, the intensities of all isotopes, or any combination thereof can be integrated.

The methods described herein are implemented using virtually any computer system and according to the following exemplary programs. Figure 1 shows an exemplary computer system. Computer system 2 includes internal and external components. The internal components include a processor 4 coupled to a memory 6. The external components include a mass-storage device 8, e.g., a hard disk drive, user input devices 10, e.g., a keyboard and a mouse, a display 12, e.g., a monitor, and usually, a network link 14 capable of connecting the computer system to other computers to allow sharing of data and processing tasks. Programs are loaded into the memory 6 of this system 2 during operation. These programs include an operating system 16, e.g., Microsoft Windows, which manages the computer system, software 18 that encodes common languages and functions to assist programs that implement the methods of this invention, and software 20 that encodes the methods of the invention in a procedural language or symbolic package. Languages that can be used to program the methods include, without limitation, Visual C/C^{++} from Microsoft. In preferred applications, the methods of the invention are programmed in mathematical software packages that allow symbolic entry of equations and high-level specification of processing, including procedures used in the execution of the programs, thereby freeing a user of the need to program procedurally individual equations or procedures. An exemplary mathematical software package useful for this purpose is Matlab from Mathworks (Natick, MA). Using the Matlab software, one can also apply the Parallel Virtual Machine (PVM) module and Message Passing Interface (MPI), which supports processing on multiple processors. This implementation of PVM and MPI with the methods herein is accomplished using methods known in the art. Alternatively, the software or a portion thereof is encoded in dedicated circuitry by methods known in the art.

Figure 2 shows a computer implemented flowchart of a method of determining an abundance of a biomolecule in a sample. In particular, Figure 2 describes determining an abundance of one biomolecule in a sample. The method, however, may also be used to determine an abundance of more than one biomolecule in the sample.

At step 100, mass spectrometry measurements are input into the program. These measurements may include MS data, MS/MS data, a plurality of scans from a separation of a biomolecule or fragment thereof, and/or structural or sequence information, e.g., obtained from searching a database. At step 102, ions measured by a mass spectrometer are assigned to a biomolecule. The biomolecule may be a fragment of a larger biomolecule. The assignation can be based on determining a mass or charge state of the ion as is described herein. The assignation can alternatively be based on sequence or structural information obtained, for example, from MS/MS and database searching as described herein. At step 104, ion counts, corresponding to intensities of ions, are integrated. Ion counts for one ion peak that are spread, for example, over a range of m/z values or scans are summed to determinc an integrated intensity for each peak. An integration of several peaks that are generated from the same biomolecule, for example, from charge states, isotopes, scans, and fragments of a larger biomolecule, or a subset thereof, may also occur. The ion counts may also be normalized with one or more internal standards. The integrated ion counts from a sample may be compared to ion counts from another sample, e.g., a reference sample, to determine an abundance of the biomolecule in the sample relative to another sample. The integrated ion counts may also be normalized with an absolute standard for that biomolecule to determine an absolute abundance of the biomolecule in the sample. The results of the analysis may be displayed to the user, e.g., on a monitor, or stored in memory. Further analysis of the data may then occur such as statistical analysis of the calculated abundances.

In one application, the invention features computer implemented modules for studying proteins. Such modules are described here as exemplars of the methods of the invention. Other biomolecules may be studied using similar modules. As is described below, the Peptide Abundance Module (PAM) determines abundances of known peptides (e.g., precursors for which MS/MS spectra have produced sequence information, or predicted peptides, such as the constituent peptides of a full-length protein or of a theoretical sequence, for which a mass and/or retention time can be predicted), and the Peptide Hunter Module (PHM) identifies and determines abundances of unknown peptides (e.g., precursors that have not yet been matched with MS/MS spectra, or for which MS/MS spectra have not yet been determined). The PHM and PAM modules of MIPS, if desired, are run simultaneously in a multiprocessing environment to reduce the time required for analysis. The multiprocessing environment, for example, includes a cluster of systems (e.g., Linux-based PCs) or servers with multiple processors (e.g., from Sun Microsystems), and the methods herein are implemented onto such distributed networks using methods known in the art (see Taylor et al. (1997) Journal of Parallel and Distributed Computing 45: 166 - 175).

The PAM and PHM offer significantly increased speed of analysis compared to performing the methods herein manually. For example, the PAM applied to finding one protein with five tryptic peptides in five samples requires approximately seven minutes to obtain an abundance of the protein. Manually, the process requires several days. In another example, the PHM for one data set requires 15 minutes to find more than 2000 proteins, but manually, the task would require more than several hundred man hours.

Peptide Abundance Module

Mass spectrometry allows the identification of a large number of peptides in a sample, for example, from MS/MS analyses. An MS/MS cycle produces peptide sequence information on a selected peptide, which may then be used to search databases comprehensively. For example, a computer is used to search available databases for a matching amino acid sequence or for a nucleotide sequence, including an expressed sequence tag (EST), whose predicted amino acid sequence matches the experimentally determined amino acid sequence. Exemplary databases useful for this purpose include, without limitation, Genbank, EMBL, NCBI, MSDB, SWISS-PROT, TrEMBL, dbEST, Human Genome Sequence database, or a user-defined database. Sequence information on compounds in the databases that contain the selected peptide may then be used to produce a list of other peptides derived from that compound using a specified cleavage technique. The mass spectra are then searched automatically for peaks corresponding to ions, e.g., from charge states or isotopes of a predicted peptide. Intensity profiles for these ions are generated through the Peptide Abundance Module (PAM). The PAM can generate intensity profiles for peptides predicted from the method of cleavage as well as those used for database searching. Peptide peaks for which MS/MS data have been acquired are matched with respect to their respective retention times.

A comprehensive list of peptide peak areas is typically generated and corrected using internal standardization, which allows the intensity of the peptides in the sample to be expressed relative to the reference peptide. Peptide peak areas for a particular peptide, for example, across bands or fractions, charge states, fragments may be combined to facilitate comparison.

Peptide peak areas from multiple peptides that are components of a compound may also be combined further to facilitate comparison. For example, the average of intensities of peptides derived from one protein may be taken as a measure for the protein abundance. A comparison of peak areas enables the identification of differences in protein abundance between experimental sets. The process is entirely automated, which facilitates data analysis.

Figure 3 shows a flowchart detailing the components of a PAM. Each component is described in detail below. A flowchart for an exemplary PAM is shown in Figure 13. This flowchart is presented for the purpose of illustrating, not limiting, the methods of the invention. An example estimate of time for PAM processing a specific sample is illustrated in Figure 14.

Data format conversion. The raw mass spectrometry data files typically consist of MS scans or a series of survey scans and MS/MS cycles for each fraction of the separation. Each mass spectrum corresponds, e.g., to an elution time period for LC or to a fraction for gel electrophoresis, or both. Each survey scan records the number of ions of each m/z value detected by the mass spectrometer. The raw mass spectrometry data files may be generated by various publicly available software packages including, without limitation, MassLynx from Micromass (Beverly, MA). To integrate MIPS with, e.g., MassLynx, software in MassLynx converts the data from the mass spectrometer, for example, into an ASCII or NetCDF format. Other software packages for obtaining mass spectrometry data have similar conversion software. Alternatively, software for data conversion is written using methods known in the art and included in the module. Optionally, data conversion, may also include merger of multiple files. File merger may also include merger of elements of the files, such as the abundances of particular precursors.

Protein/Gene Identification. The raw mass spectrometry data is submitted for compound, e.g., protein, identification using a tool such as Mascot from Matrix Science (London, United Kingdom), ProteinLynx Global Server from Micromass SEQUEST/TurboSEQUEST from Thermo Finnigan

(San Jose, CA), or Sonar MS/MS from ProteoMetrics (New York, NY). This analysis generates a list of proteins that are likely to exist in the sample under analysis.

Peptide List Generation. In this component a list of annotated theoretical peptide masses are generated from the list of identified proteins, identified genes, and the raw mass spectrometry data as follows. Cleavage (e.g., tryptic) peptides of the identified compounds may be predicted, and theoretical masses of these peptides may be generated. The mass of a peptide can then be estimated from its amino acid sequence. The mass may also be based upon the mass of a matched peptide from a sample. Other peptide masses of interest can be added to the list of theoretical peptide masses. For example, the mass of a post-translationally modified peptide (e.g., a phosphorylated or glycosylated peptide) is estimated from the unmodified mass. Adding these masses to the list allows MIPS to track the relative abundance of modified to unmodified peptides. The theoretical peptides masses may be annotated by their source (e.g., protein, gene, raw data, or modified peptide).

For normalization across samples, one or more internal peptide standards are optionally added into each sample. The theoretical peptide mass of any internal standards are then added to the theoretical peptide mass list.

Obtain Observed Peptide Masses. The peptide masses measured by the spectrometer may differ from the theoretical peptide masses depending on the accuracy of mass measurement as defined by the instrument parameters. This discrepancy can be corrected by determining the mass with the maximum ion count within a predefined range around the theoretical peptide mass according to known correction methods. This range is defined by the accuracy of the mass spectrometry instrument and also the user defined tolerance for false positives and false negatives. A list of observed peptide masses is thus generated from the list of theoretical peptide masses.

Integration of Observed Peptide Ion Counts Over Scans, Isotopes, and Charge States. To obtain the abundance of a peptide, the intensities of all ions, or a subset thereof, for that peptide are counted. In addition, the ions of a peptide that occur in different spectra because of the separation, if desired, are integrated.

Each peptide in the sample under analysis is eluted over a series of scans within the raw mass spectrometry data (see Figure 4). This collection of scans is referred to as the "scan window" of the peptide. Each scan in this window corresponds to a fraction from the separation, e.g., an elution time from LC. For LC, a retention time for a peptide is determined based on the intensity of an ion corresponding to that peptide as a function of scan number, which corresponds to time.

In each of the scans within a peptide's scan window, the peptide occurs in multiple charge states, typically +1, +2, +3, or +4. Scans record mass/charge (m/z). For example, the ion peak for a doubly-protonated peptide (charge +2) is $(1974 + (2 \times 1.0078))/2 = 988.0078$ when the uncharged mass is 1974 (see Figure 5) and when the peptide is ionized by electrospray ionization. For each charge state of a peptide in a scan, the ions of the peptide are distributed over multiple isotopes. These peptide isotopes are the result of the natural abundances of isotopes of the constituent atoms. Five or more isotopes per charge state may be present. The m/z of these isotopes is predicted from the mass and charge of the peptide. For example, a peptide with mass 1974 and a charge of +3 has a m/z of 659 (see Figure 6).

For each isotope examined, an integration window is defined, over which ion counts are summed. This window is typically required because of the resolution of the mass spectrometer and the accuracy in mass measurement. The width of this window depends, for example, on the peak width at a predefined height for an isotopic peak of the peptide. For a Gaussian peak, the ratio of peak width at half height (PW_{50}) and peak width at 5% height (PW_5) should be constant and equal to approximately 2.2. The relationship linking

PW_5 to the m/z value is represented by $PW_5 = (2.2 \times m/z)/RES$ where RES is the instrumental resolution as defined at PW_{50}. Once the PW_5 is calculated for a peak, ion counts from $m/z - PW_5/2$ to $m/z + PW_5/2$ are summed, for example, by adding the measured intensities for data points falling within the window, to produce the integrated intensity of that peak.

The procedure for determining the abundance of a peptide with observed mass is to sum the ion counts over isotopes of all charge states, or a subset thereof, in scans where the sum of the ion counts are above a predefined signal to noise threshold.

The retention time of each peptide is also obtained in this process. If two peptides of the same mass but significantly different retention times are detected, then two separate entries for these peptides are created since they are likely different peptides that happen to have the same mass. This duplicity may be noted in the data.

Normalization. The abundances of the observed peptide masses may be corrected for instrument variation using the abundance of one or more internal standards, e.g., Leu-enkephalin. A standard may be added from an external source (e.g., a synthetic peptide standard) or may be intrinsic to the sample (e.g., peptides derived from a protein marker believed to remain constant across conditions to be compared). Preferably, several such standards are used across the range of retention times. An absolute abundance of a peptide may be determined using intensity data from purified or synthetic versions of that peptide. Optionally, a sub-component of the module can be used to acquire data about internal standards, including, but not limited to, m/z values, intensity, and retention times. Such data may be used in combination with known reference values to correct and/or predict the corresponding values for the standards themselves, as well as of other peptides and precursors. For example, data derived from internal standards (such as, RT and m/z values) as well RT from original MS-MS peptide data can be used to predict the retention times (RT) of peptides in MS data from the LC-MS survey. Data about internal standards and peptides with m/z values and corrected retention times may be produced for a subset or a in file-wide data integration / correction.

Integration Over Fractions or Bands. If samples analyzed by mass spectrometry are excised from 1D gels, the abundance of an observed peptide is usually integrated over neighboring bands since the peptide can appear in several bands. The same peptide in neighboring bands is then identified, e.g., by mass, retention time, and MS/MS. If samples are analyzed by multidimensional LC (e.g., 2D), the abundance is typically integrated over salt fractions.

Protein/Gene Abundance Statistical Analyses. The abundance of a protein or gene is proportional to the abundance of its peptide abundances. Statistical analyses may be performed to determine the relative abundance of a protein or gene across samples. For example, the distribution of peptide abundances in normal tissue is compared to the distribution of the peptide abundances in diseased tissue using a statistical test of significance. In addition, the relative abundance of post-translationally modified and unmodified proteins within a sample may be determined from the relative abundance of modified to unmodified peptides. Abundances of isoforms of a protein may also be determined. Similarly, the relative abundance of splice variants of a gene may be determined.

Individual Peptide Abundance Statistical Analyses. Observed peptides not assigned to a protein or gene through protein identification may also be compared across samples. Significant differences in abundance across samples indicate interesting peptides for further analysis.

Those peptides that occur in the raw mass spectrometry data but not within any of the identified proteins or genes may be obtained, e.g., from the raw mass spectrometry data, e.g., using pattern detection methods.

Peptide Hunter Module

The Peptide Hunter Module (PHM) differs from the Peptide Abundance Module in that proteins t and their corresponding tryptic peptides are "precursors" and need not have been identified, such as by being sequenced by MS-MS, prior to using PHM. The PHM therefore mines raw mass spectrometry data for peptides, calculates their abundance (see Figures 16 and 17 for an example), and may render them comparable between samples in part by correction of their mass, intensity, and retention times through the use of internal standards. PHM allows the identification of full-length proteins a peptide may be from through peptide mass fingerprinting. PHM also permits the generation of a list of precursors for further identification by a round of MS/MS. Figure 7 is a flowchart detailing the components of a Peptide Hunter Module (PHM). Solid rectangles represent processing components of MIPS, dashed rectangles represent processing components that are not within MIPS and entries without a rectangle are data files. A flowchart for an exemplary PHM is shown in Figure 15. Each component is described in detail below. This flowchart is presented for the purpose of illustrating, not limiting, the methods of the invention.

Data Format Conversion. Data format conversion occurs as is described above in the PAM.

Determination of a Threshold. Since the PHM mines the survey scans in the raw mass spectrometry data for evidence of peptides, a threshold of ion intensity is defined to differentiate signal from peptide ions from those of noise. This threshold is estimated for all scans by using methods known in the arts, such methods include, without limitation, the method of Maximum Entropy.

Find Charge States of Peptides in Survey Scans. A survey scan of raw mass spectrometry data is searched for evidence of charged states of peptides. The pattern of a charge state of a peptide is depicted in Figure 6. Each charge state consists of a pattern of isotopic peaks. The isotopes of the charged state are separated in a spectrum by 1.0034/z, where z is the charge of the peptide. The "first isotope" of a charge state can be located at a specific m/z value with an isotope located at ((m/z value) + 1.0034/z), but without an isotope located at ((m/z value) - 1.0034/z) in the spectrum. The second isotope can be located at ((m/z value) + 1.0034/z) in the spectrum, and so on.

To identify a charge state for a peptide, a data point corresponding to an m/z can be selected, e.g., on the basis of intensity, from the data in a spectrum. The data can then be searched systematically for neighboring peaks separated by 1.0034/z for a defined number of charges, e.g., +4, +3, +2, and +1. The program searches an appropriate region around 1/z to compensate for uncertainty in the experimental data. The charges can be searched in order from highest to lowest until a peak is found. This order is typically required since, for example, a +4 charged peptide could be mistakenly interpreted as a +1 charged peptide since the +4 charged peptide and the +1 charged peptide both have isotopes at (m/z value of first isotope + 1). If no neighboring peaks are found, a charge state cannot be assigned using this method. If a neighboring peak is present, for example, at m/z + 0.33, then the charge state can be identified by the separation, which in this case corresponds to the +3 state (Figure 6). Isotopes in a charge state are identified based on one peak and the separation (1.0034/z). Isotopes of a charge state may be assigned to the same mass or m/z, e.g., the mass or m/z of the first isotope, to facilitate integration of peaks originating from the same peptide. The search may require that a peak be a first isotope, and that the second isotope be at least a specified fraction (possibly greater than 1) of the first isotope. Once a charge state is identified, a mass of the peptide may be calculated and used to search for other charge states from the same peptide. By using this procedure, many peaks may be identified from the initial identification of one peak.

In one embodiment, for each peak, m, in the scan, beginning with the most intense peak and progressing to the least intense peak with intensities above the threshold, t, the following steps occur. Alternatively, only a selected number are analyzed as follows. Ion counts within a window, w, around data point m are integrated to obtain abundance, A1. Ion counts within a window, w, around m + 0.25 are then integrated to obtain abundance, A2. Ion counts within a window, w, around m - 0.25 are then integrated to obtain abundance, A0. If A2 is greater than p × A1 and A1 is greater than q × A0, then m is the first isotope of the +4 charge state of a peptide. Otherwise, repeat the above steps with 0.25 replaced with 0.33, 0.5, and 1 to test for the +3, +2, and +1 charge states. The parameters w, t, p, and q are user defined. The threshold ensures that only peaks of sufficient intensity are examined. The parameters p and q can ensure that peak m is a first isotope by requiring that the second isotope be at least a defined fraction of the first isotope, and that another isotope is not present at ((m/z value) – 1/z). Redundancy in the form of multiply identified peptides may be eliminated.

Determine Uncharged Peptide Masses. A peptide can occur in many charge states in the scans of the raw mass spectrometry data, and all or a portion of these charge states may be collected for the peptide. Charged peptides in a scan are assigned to an uncharged peptide mass using the formula P=(m/z × z) – (1.0078 × z), where P is the uncharged mass, m/z is measured by the spectrometer, and z is the charge for electrospray ionization. Other ionization schemes are known in art, and the formula is modified accordingly. Software used in the PHM may also require that peptides assigned to an uncharged peptide mass have similar retention times. In the example of Figure 6, the PHM would detect a +3 charged peptide with an uncharged mass, P = (658.96 × 3) – (1.0078 × 3) = 1973.86.

Integrate Abundances Over Scans, Charge States, and Isotopes. Integrated ion counts for individual peaks may be calculated as is described in the PAM above. For each of the uncharged peptide masses, a measure of the abundance of the uncharged peptide mass may be obtained by an integration of ion counts, e.g., over scans, charge states, and isotopes, or combinations thereof. The abundance need not take into account the abundance of every ion count for a particular peptide, though a consistent subset of the possibilities, e.g., just the +2 and +3 charge states, should be used throughout the sample. For each uncharged peptide, the previous step predetermines those charge states and isotopes that correspond to the same uncharged mass. The retention time of the uncharged peptide mass can be predicted from the scan window as in the PAM described above.

Normalization by One or More Internal Standards. The abundances of the observed peptide masses are optionally corrected for instrument variation using the abundance of one or more internal standards according to standard methods, and may be performed similarly as for PAM above. If the PHM is executed after the PAM, then a second normalization is unnecessary, although it is desirably for it to be performed for ions for which no MS-MS was obtained.

Integration Over Fractions or Bands. If samples analyzed by mass spectrometry are excised from 1D gels, the abundance of an observed peptide is typically integrated over neighboring bands since the peptide might appear in several bands. The same peptide in neighboring bands is identified, e.g., by mass, retention time, and MS/MS. If samples are analyzed by multidimensional LC (e.g., 2D), the abundance is typically integrated over salt fractions.

Individual Peptide Abundance Statistical Analyses. The list of peptides masses, their abundances, and retention times are used for various analyses, such as protein identification by mass fingerprinting; protein identification, through defining peptides for a further round of MS/MS; protein identification that combines matching MS/MS and mass fingerprinting, which can increase the peptide coverage of a protein and assist in differentiating between similar proteins in a family or between splice variants and between polymorphisms; and determining low abundance peptides present in the raw mass spectrometry data, which may correspond to low abundance proteins in the sample being analyzed.

Expression Profiling

The methods of the present invention can be used to determine the relative abundance of a biomolecule or fragment thereof, e.g., proteins, in samples (see Figures 18 and 19). Samples being analyzed are compared to a reference sample, or samples. This comparison, or expression profile, is used, e.g., to determine if biomolecules, e.g., proteins, are present in abnormally high or low amounts compared to the reference. The determination of a difference in expression of a species in a sample relative to a reference sample is used, e.g., to diagnose disease in a patient, to determine natural variance in a population, or to determine the genotype of an individual. A comparison of protein abundances between normal and tumor cells for an individual (see Figure 18), or across a population of patients, would be exemplary applications.

Drug Targets

Once a protein is identified in a public or private database, the gene encoding the protein is cloned and introduced into bacterial, yeast, or mammalian host cells. Where such a gene is not identified in a database, the gene encoding the protein is cloned, using a degenerate set of probes that encode an amino acid sequence of the protein as determined by the methods discussed above. Where a database contains one or more partial nucleotide sequences that encode an experimentally determined amino acid sequence of the protein, such partial nucleotide sequences (or their complement) serve as probes for cloning the gene, obviating the need to use degenerate sets.

Cells genetically engineered to express such a recombinant protein can be used in a screening program to identify other proteins or drugs that specifically interact with the recombinant protein, or to produce large quantities of the recombinant protein, e.g. for therapeutic administration.

In addition, a protein identified according to the present invention can be used to generate antibodies, for example, by administering the protein to an animal, such as a mouse, rat, or rabbit, for production of polyclonal or monoclonal antibodies using standard methods known in the art. Such antibodies are useful in diagnostic and prognostic tests and for purification of large quantities of the protein, for example, by antibody affinity chromatography. Antibodies may also be used for immunotherapy, such as might be used in the treatment of cancer.

EXAMPLES

These aforementioned methods, and the reagents and techniques for carrying out these steps, are now described in detail using particular examples. The examples are provided for the purpose of illustrating the invention and should not be construed as limiting.

Example 1

Reproducibility

A total cellular lysate of U937 cells was prepared. Fifty µg of this lysate was mixed with 30 ng of bovine serum albumin (BSA). Five samples of this mixture were prepared, and each sample was separated by SDS gel electrophoresis. A sixth sample was prepared without the addition of BSA. After electrophoresis, the gel was stained, and the band containing the BSA was excised from the gel.

Tryptic digestion of the BSA band was performed according to standard methods. Three hundred and fifty fmol of Leu-enkephalin was added to each tryptic digest as an internal standard. The peptide mixture was separated using a Waters CapLC HPLC system that was coupled to a Micromass quadropole-time of flight (Q-ToF) mass spectrometer. The conditions for the separation of this mixture were as follows: a reversed phase column (1D, 10 cm ×75 µm, C18), a flow rate of 300 nL/min, and a linear gradient of 10% to 80%

acetonitrile/deionized water (containing 0.2% formic acid) in 25 minutes. The settings of the mass spectrometer were as follows: MS scan using a mass range from m/z 400 to 1500, a solvent delay of 5 minutes, a scan time of 19 minutes, a scan rate of 1 second, and an interscan delay of 0.1 second.

A theoretical list of BSA-derived peptides was generated from the known protein sequence of BSA and the known cleavage preference of trypsin. A subset of these peptides, which were detected in the mass spectra, was chosen. This list, together with the mass spectrometer data files, was entered into the PAM to generate a list of peptide abundances. The average peptide abundance was calculated as a measure of protein abundance. The average and standard deviation from the 5 replicates were calculated. For all five samples, the measured BSA abundance was close to the average, with a relative standard deviation of approximately 2 to 3 %. Table 1 shows the relative abundance of BSA, human dihydropyrimidinase-related protein 2 (DHP), heat shock protein 89 (HSP89), and inosine monophosphate dehydrogenase (IMP) as determined by the PAM for five samples. DHP, HSP89, and IMP co-migrated with BSA. This experiment established that the automated analysis yields highly reproducible results when the same sample is analyzed several times.

Table 1

Sample Number	BSA RSD 2.5%	DHP RSD 7.6%	HSP89 RSD 2.8%	IMP 11.6%
1	451	1860	1011	1795
2	466	1596	977	1769
3	444	1583	1028	2125
4	456	1622	975	1923
5	474	1801	1030	2308

Example 2

Linearity

U937 cells were cultured for 48 hours with or without 25 nM phorbol myristate acetate (PMA) to generate macrophages and untransformed cells (monocytes) using standard culturing techniques. Approximately 100 million cells suspended in 10 mM Tris / 200 mM sucrose, pH 7.5 homogenization buffer were placed into a cavitations chamber, pressurized with N_2 at 1000 psi and kept on ice for 60 minutes. After each incubation period, samples were released to atmospheric pressure rapidly, and centrifuged at 900 g for 15 minutes at 4 °C (low speed centrifugation). The supernatant (post nuclear supernatant, PNS) was collected and either kept or immediately centrifuged at 40000 rpm for 60 minutes at 4 °C to produce the post nuclear membrane (PNM) samples.

Five samples of 50 μg of monocyte or macrophage proteins were prepared. The first sample contained only monocyte proteins, the second sample contained 75 % monocyte and 25 % macrophage proteins, the third sample contained 50 % of both monocyte and macrophage protein, the fourth sample contained 25 % monocyte and 75 % macrophage protein, and the fifth sample contained only macrophage proteins. To each sample, 50 ng of BSA was added. These samples were separated by electrophoresis, and the gel was stained using standard methods.

From each lane, 10 bands were excised (with BSA being located approximately in the middle of the lane). Each band was analyzed by mass spectrometry as described in the previous Example. The tryptic peptide mixture was separated using a Waters CapLC HPLC system coupled to a Micromass Q-TOF mass spectrometer. The conditions for the separation of this mixture were as follows: a reversed phase column (1D, 10 cm × 75 μm, C18), a flow rate of 300 nL/min and a linear gradient of 5% to 60% acetonitrile/deionized water (containing 0.2% formic acid) in 15 minutes. The mass spectrometer settings were as follows: MS and MS/MS scans using a mass range from m/z 400 to 1600, a scan time of 24 minutes, a scan rate of 1 second, and an interscan delay of 0.1 second.

Figure 8 shows mass spectrometry data obtained for the HSP90 protein. Three peptides were identified and characterized by their LC retention times, a MS survey scan, and MS/MS. Figure 9 shows a table of the relative abundance of each of the peptides as a function of the concentration of macrophage proteins and monocyte proteins. The HSP90 is expressed to a greater extent in monocytes than in macrophage. Figure 12A shows the linearity of the expression of HSP90.

Figure 10 shows mass spectrometry data obtained for the mutant desmin protein. Three peptides were identified and characterized by their LC retention times, a MS survey scan, and MS/MS. Figure 11 shows a table of the relative abundance of each of the peptides as a function of the concentration of macrophage proteins and monocyte proteins. The mutant desmin is expressed to a greater extent in macrophage than in monocytes. Figure 12B shows the linearity of the expression of mutant desmin.

Other Embodiments

All patents, patent applications, and publications referenced herein are hereby incorporated by reference.

Other embodiments are within the following claims.

What is claimed is:

CLAIMS

1. A method for determining abundance of a biomolecule in a biological sample, said method comprising the steps of:

a) providing a biological sample comprising a plurality of biomolecules;

b) generating a plurality of ions of said biomolecules;

c) performing mass spectrometry measurements on the plurality of ions, thereby obtaining ion counts for the biomolecules;

d) assigning an ion to a biomolecule; and

e) integrating the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

2. The method of claim 1, wherein said biomolecule is a protein.

3. The method of claim 2, wherein said protein is from an isolated organelle.

4. The method of claim 3, wherein said organelle is selected from the group consisting of mitochondria, chloroplasts, ER, Golgi, endosomes, lysosomes, phagosomes, peroxisomes, secretory vesicles, transport vesicles, nuclei, and plasma membrane.

5. The method of claim 2, wherein the protein is a cytosolic or cytoskeletal protein.

6. The method of claim 1, wherein said biomolecule is unlabeled.

7. The method of claim 1, wherein said biomolecule is underivatized.

8. The method of claim 1, wherein said biomolecule is a cleaved biomolecule.

9. The method of claim 8, wherein said cleaved biomolecule is unlabeled.

10. The method of claim 8, wherein said cleaved biomolecule is underivatized.

11. The method of claim 8, wherein said biomolecule is cleaved with an enzyme.

12. The method of claim 11, wherein said enzyme is trypsin.

13. The method of claim 1, further comprising separating the plurality of biomolecules prior to step (b).

14. The method of claim 13, wherein separation is carried out by chromatography, electrophoresis, immunoisolation, or centrifugation.

15. The method of claim 13, wherein said biological sample includes one or more internal standards and wherein the retention time of an ion is corrected using said one or more internal standards.

16. The method of claim 1, further comprising assaying a second biological sample.

17. The method of claim 1, wherein said biological sample includes one or more internal standards.

18. The method of claim 1, where step (c) further comprises determining structural or sequence information of an ion of the biomolecule.

19. The method of claim 18, wherein structural or sequence information is obtained from MS/MS.

20. The method of claim 19, wherein a list of one or more biomolecules is provided to select an ion of a biomolecule for MS/MS analysis.

21. The method, of claim 20, wherein said list is an inclusion list.

22. The method of claim 20, wherein said list is an exclusion list.

23. The method of claim 18, further comprising using a computer procedure in step (d) to identify a biomolecule comprising the structure or sequence of the ion from a database.

24. The method of claim 23, wherein said computer procedure is selected from the group consisting of Mascot®, Protein Lynx Global Server, SEQUEST®/TurboSEQUEST, PepSEQ, SpectrumMill, or Sonar MS/MS.

25. The method of claim 23, wherein said database is the Genbank®, EMBL, NCBI, MSDB, SWISS-PROT®, TrEMBL, dbEST, or Human Genome Sequence database.

26. The method of claim 23, wherein step (d) is carried out using a computer procedure that assigns the ion to the biomolecule identified ~~from~~ said database.

27. The method of claim 26, wherein step (d) is carried out using a computer procedure that assigns the ion to the biomolecule by calculating an uncharged mass for the ion.

28. The method of claim 1, wherein step (d) is carried out using peptide mass fingerprinting.

29. The method of claim 1, wherein step (e) is carried out using a computer procedure that integrates ion counts of at least two ions corresponding to the biomolecule.

30. The method of claim 29, wherein the integration is over one or more charge states, isotopes, scans, fragments of the biomolecule, fractions of a separation, or a combination thereof.

31. The method of claim 1, wherein said method further comprises calculating an abundance of the biomolecule relative to a control biological sample.

32. The method of claim 1, wherein said method further comprises calculating abundances of a plurality of the biomolecules relative to a control biological sample.

33. The method of claim 31, wherein the abundance is used to diagnose a disease or condition.

34. The method of claim 31, wherein abundance is used to determine a biomolecule to target with a drug.

35. The method of claim 31, wherein an increase or decrease in abundance or the presence or absence of a biomolecule in the biological sample relative to the control biological sample is indicative of a disease or condition.

36. The method of claim 31, wherein the abundance is used to determine an amount of an isoform of a biomolecule.

37. A computer implemented method for determining abundance of a biomolecule in a biological sample, said method comprising the steps of:

a) inputting mass spectrometry data comprising ion counts for a plurality of biomolecules into a computer;

b) assigning an ion to a biomolecule; and

c) integrating the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

38. The computer implemented method of claim 26, wherein said biomolecule is a protein.

39. The computer implemented method of claim 38, wherein said protein is from an isolated organelle.

40. The computer implemented method of claim 39, wherein said organelle is selected from the group consisting of mitochondria, chloroplasts, ER, Golgi, endosomes, lysosomes, phagosomes, peroxisomes, secretory vesicles, transport vesicles, nuclei, and plasma membrane.

41. The computer implemented method of claim 38, wherein the protein is a cytosolic or cytoskeletal protein.

42. The computer implemented method of claim 37, wherein said biomolecule is unlabeled.

43. The computer implemented method of claim 37, wherein said biomolecule is underivatized.

44. The computer implemented method of claim 37, wherein said biomolecule is a cleaved biomolecule.

45. The computer implemented method of claim 44, wherein said cleaved biomolecule is unlabeled.

46. The computer implemented method of claim 44, wherein said cleaved biomolecule is underivatized.

47. The computer implemented method of claim 44, wherein said biomolecule is cleaved with an enzyme.

48. The computer implemented method of claim 47, wherein said enzyme is trypsin.

49. The computer implemented method of claim 37, wherein the plurality of biomolecules is separated prior to the acquisition of mass spectrometry data.

50. The computer implemented method of claim 49, wherein separation is carried out by chromatography, electrophoresis, immunoisolation, or centrifugation.

51. The computer implemented method of claim 49, wherein said biological sample includes one or more internal standards and wherein the retention time of an ion is corrected using said one or more internal standards.

52. The computer implemented method of claim 37, further comprising assaying a second biological sample.

53. The computer implemented method of claim 37, wherein said biological sample includes one or more internal standards.

54. The computer implemented method of claim 37, where the mass spectrometry data further comprises structural or sequence information of an ion of the biomolecule.

55. The computer implemented method of claim 54, wherein said structural or sequence information is obtained from MS/MS.

56. The method of claim 55, wherein a list of one or more biomolecules is provided to select an ion of a biomolecule for MS/MS analysis.

57. The method, of claim 55, wherein said list is an inclusion list.

58. The method of claim 55, wherein said list is an exclusion list.

59. The computer implemented method of claim 54, further comprising using the structural or sequence information to identify a biomolecule from a database.

60. The computer implemented method of claim 59, wherein the biomolecule is identified using a computer procedure selected from the group consisting of Mascot®, Protein Lynx Global Server, SEQUEST®/TurboSEQUEST, PepSEQ, SpectrumMill, or Sonar MS/MS.

61. The computer implemented method of claim 59, wherein said database is the Genbank®, EMBL, NCBI, MSDB, SWISS-PROT®, TrEMBL, dbEST, or Human Genome Sequence database.

62. The computer implemented method of claim 59, wherein in step (b) the ion is assigned to the biomolecule identified from said database.

63. The computer implemented method of claim 37, wherein in step (b) the ion is assigned to the biomolecule by calculating an uncharged mass for the ion.

64. The computer implemented method of claim 37, wherein in step (b) the ion is assigned to the biomolecule by peptide mass fingerprinting.

65. The computer implemented method of claim 37, wherein in step (c) the integration is over one or more charge states, isotopes, scans, fragments of the biomolecule, fractions of a separation, or a combination thereof.

66. The computer implemented method of claim 37, wherein said method further comprises calculating an abundance of the biomolecule relative to a control biological sample.

67. The computer implemented method of claim 37, wherein said method further comprises calculating abundances of a plurality of biomolecules relative to a control biological sample.

68. The computer implemented method of claim 66, wherein the abundance is used to diagnose a disease or condition.

69. The computer implemented method of claim 66, wherein abundance is used to determine a biomolecule to target with a drug.

70. The computer implemented method of claim 66, wherein an increase or decrease in abundance or the presence or absence of a biomolecule in the biological sample relative to the control biological sample is indicative of a disease or condition.

71. The computer implemented method of claim 66, wherein the abundance is used to determine an amount of an isoform of a biomolecule.

72. A computer-readable memory having stored thereon a program for determining abundance of a biomolecule in a biological sample comprising:

a) computer code that receives as input mass spectrometry data comprising ion counts for a plurality of biomolecules;

b) computer code that assigns an ion to a biomolecule; and

c) computer code that integrates the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

73. The computer-readable memory of claim 72, wherein said biomolecule is a protein.

74. The computer-readable memory of claim 73, wherein said protein is from an isolated organelle.

75. The computer-readable memory of claim 74, wherein said organelle is selected from the group consisting of mitochondria, chloroplasts, ER, Golgi, endosomes, lysosomes, phagosomes, peroxisomes, secretory vesicles, transport vesicles, nuclei, and plasma membrane.

76. The computer-readable memory of claim 73, wherein the protein is a cytosolic or cytoskeletal protein.

77. The computer-readable memory of claim 72, wherein said biomolecule is unlabeled.

78. The computer-readable memory of claim 72, wherein said biomolecule is underivatized.

79. The computer-readable memory of claim 72, wherein said biomolecule is a cleaved biomolecule.

80. The computer-readable memory of claim 79, wherein said cleaved biomolecule is unlabeled.

81. The computer-readable memory of claim 79, wherein said cleaved biomolecule is underivatized.

82. The computer-readable memory of claim 79, wherein said biomolecule is cleaved with an enzyme.

83. The computer-readable memory of claim 82, wherein said enzyme is trypsin.

84. The computer-readable memory of claim 72, wherein the plurality of biomolecules is separated prior to the acquisition of mass spectrometry data.

85. The computer-readable memory of claim 84, wherein separation is carried out by chromatography, electrophoresis, immunoisolation, or centrifugation.

86. The computer-readable memory of claim 84, wherein said biological sample includes one or more internal standards and wherein the retention time of an ion is corrected using said one or more internal standards.

87. The computer-readable memory of claim 72, further comprising assaying a second biological sample.

88. The computer-readable memory of claim 72, wherein said biological sample includes one or more internal standards.

89. The computer-readable memory of claim 72, where the mass spectrometry data further comprises structural or sequence information of an ion of the biomolecule.

90. The method of claim 89, wherein said structural or sequence information is obtained from MS/MS.

91. The method of claim 90, wherein a list of one or more biomolecules is provided to select an ion of a biomolecule for MS/MS analysis.

92. The method, of claim 91, wherein said list is an inclusion list.

93. The method of claim 91, wherein said list is an exclusion list.

94. The computer-readable memory of claim 89, wherein in step (b) the structural or sequence information is used to identify a biomolecule from a database.

95. The computer-readable memory of claim 94, wherein the biomolecule is identified using a computer procedure selected from the group consisting of Mascot®, Protein Lynx Global Server, SEQUEST®/TurboSEQUEST, PepSEQ, SpectrumMill, or Sonar MS/MS.

96. The computer-readable memory of claim 94, wherein said database is the Genbank®, EMBL, NCBI, MSDB, SWISS-PROT®, TrEMBL, dbEST, or Human Genome Sequence database.

97. The computer-readable memory of claim 94, wherein in step (b) the ion is assigned to the biomolecule identified from said database.

98. The computer-readable memory of claim 72, wherein in step (b) the ion is assigned to the biomolecule by calculating an uncharged mass for the ion.

99. The computer-readable memory of claim 72, wherein in step (b) the ion is assigned to the biomolecule by peptide mass fingerprinting.

100. The computer-readable memory of claim 72, wherein in step (c) the integration is over one or more charge states, isotopes, scans, fragments of the biomolecule, fractions of a separation, or a combination thereof.

101. The computer-readable memory of claim 72, wherein the computer code in step (c) further calculates the abundance of the biomolecule relative to a control biological sample.

102. The computer-readable memory of claim 72, wherein the computer code in step (c) further calculate abundances of a plurality of biomolecules relative to a control biological sample.

103. The computer-readable memory of claim 101, wherein the abundance is used to diagnose a disease or condition.

104. The computer-readable memory of claim 101, wherein abundance is used to determine a biomolecule to target with a drug.

105. The computer-readable memory of claim 101, wherein an increase or decrease in abundance or the presence or absence of a biomolecule in the biological sample relative to the control biological sample is indicative of a disease or condition.

106. The computer-readable memory of claim 101, wherein the abundance is used to determine an amount of an isoform of a biomolecule.

107. A system for determining abundance of a biomolecule in a biological sample comprising:

a) a mass spectrometry data input module that receives data comprising ion counts for a plurality of biomolecules;

b) an ion assigning module responsive to the data input module, wherein said ion assigning module assigns an ion to a biomolecule; and

c) an ion integrating module responsive to the ion assigning module, wherein said ion integrating module integrates ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample.

108. The system of claim 74, wherein said system comprises a processor and a memory coupled to said processor, said memory encoding said data input module, said ion assigning module, and said ion integrating module.

109. The system of claim 107, wherein said biomolecule is a protein.

110. The system of claim 109 wherein said protein is from an isolated organelle.

111. The system of claim 110 wherein said organelle is selected from the group consisting of mitochondria, chloroplasts, ER, Golgi, endosomes, lysosomes, phagosomes, peroxisomes, secretory vesicles, transport vesicles, nuclei, and plasma membrane.

112. The system of claim 109, wherein the protein is a cytosolic or cytoskeletal protein.

113. The system of claim 107, wherein said biomolecule is unlabeled.

114. The system of claim 107, wherein said biomolecule is underivatized.

115. The system of claim 107, wherein said biomolecule is a cleaved biomolecule.

116. The system of claim 115, wherein said cleaved biomolecule is unlabeled.

117. The system of claim 115, wherein said cleaved biomolecule is underivatized.

118. The system of claim 115, wherein said biomolecule is cleaved with an enzyme.

119. The system of claim 118, wherein said enzyme is trypsin.

120. The system of claim 107, wherein the plurality of biomolecules is separated prior to the acquisition of mass spectrometry data.

121. The system of claim 120, wherein separation is carried out by chromatography, electrophoresis, immunoisolation, or centrifugation.

122. The system of claim 120, wherein said biological sample includes one or more internal standards and wherein the retention time of an ion is corrected using said one or more internal standards.

123. The system of claim 107, further comprising assaying a second biological sample.

124. The system of claim 107, wherein said biological sample includes one or more internal standards.

125. The system of claim 107, where the mass spectrometry data further comprises structural or sequence information of an ion of the biomolecule.

126. The system of claim 125, wherein said structural or sequence information is obtained from MS/MS.

127. The method of claim 126, wherein a list of one or more biomolecules is provided to select an ion of a biomolecule for MS/MS analysis.

128. The method, of claim 127, wherein said list is an inclusion list.

129. The method of claim 127, wherein said list is an exclusion list.

130. The system of claim 125, wherein the structural or sequence information is used to identify a biomolecule from a database.

131. The system of claim 130, wherein the biomolecule is identified using a computer procedure selected from the group consisting of Mascot®, Protein Lynx Global Server, SEQUEST®/TurboSEQUEST, PepSEQ, SpectrumMill, or Sonar MS/MS.

132. The system of claim 130, wherein said database is the Genbank®, EMBL, NCBI, MSDB, SWISS-PROT®, TrEMBL, dbEST, or Human Genome Sequence database.

133. The system of claim 130, wherein in step (b) the ion is assigned to the biomolecule identified from said database.

134. The system of claim 107, wherein in step (b) the ion is assigned to the biomolecule by calculating an uncharged mass for the ion.

135. The system of claim 107, wherein in step (b) the ion is assigned to the biomolecule by peptide mass fingerprinting.

136. The system of claim 107, wherein in step (c) the integration is over one or more charge states, isotopes, scans, fragments of the biomolecule, fractions of a separation, or a combination thereof.

137. The system of claim 107, wherein said method further comprises calculating an abundance of the biomolecule relative to a control biological sample.

138. The system of claim 107, wherein said method further comprises calculating abundances of a plurality of the biomolecules relative to a control biological sample.

139. The system of claim 137, wherein the abundance is used to diagnose a disease or condition.

140. The system of claim 137, wherein abundance is used to determine a biomolecule to target with a drug.

141. The system of claim 137, wherein an increase or decrease in abundance or the presence or absence of a biomolecule in the biological sample relative to the control biological sample is indicative of a disease or condition.

142. The system of claim 137, wherein the abundance is used to determine an amount of an isoform of a biomolecule.

143. A method for displaying information on abundance of a biomolecule in a biological sample to a user, said method comprising the steps of:

a) inputting mass spectrometry data comprising ion counts for a plurality of biomolecules into a computer;

b) assigning an ion to a biomolecule;

c) integrating the ion counts of the biomolecule, thereby determining the abundance of the biomolecule in the biological sample; and

d) displaying the abundance of the biomolecule.

144. The method of claim 143, wherein step (d) further comprises storing the abundance of the biomolecule in a memory.

145. The method of claim 143, wherein said biomolecule is a protein.

146. The method of claim 145, wherein said protein is from an isolated organelle.

147. The method of claim 146, wherein said organelle is selected from the group consisting of mitochondria, chloroplasts, ER, Golgi, endosomes, lysosomes, phagosomes, peroxisomes, secretory vesicles, transport vesicles, nuclei, and plasma membrane.

148. The method of claim 145, wherein the protein is a cytosolic or cytoskeletal protein.

149. The method of claim 143, wherein said biomolecule is unlabeled.

150. The method of claim 143, wherein said biomolecule is underivatized.

151. The method of claim 142, wherein said biomolecule is a cleaved biomolecule.

152. The method of claim 151, wherein said cleaved biomolecule is unlabeled.

153. The method of claim 151, wherein said cleaved biomolecule is underivatized.

154. The method of claim 151, wherein said biomolecule is cleaved with an enzyme.

155. The method of claim 154, wherein said enzyme is trypsin.

156. The method of claim 143, wherein the plurality of biomolecules is separated prior to the acquisition of mass spectrometry data.

157. The method of claim 156, wherein separation is carried out by chromatography, electrophoresis, immunoisolation, or centrifugation.

158. The system of claim 156, wherein said biological sample includes one or more internal standards and wherein the retention time of an ion is corrected using said one or more internal standards.

159. The method of claim 143, further comprising assaying a second biological sample.

160. The method of claim 143, wherein said biological sample includes one or more internal standards.

161. The method of claim 143, where the mass spectrometry data further comprises structural or sequence information of an ion of the biomolecule.

162. The method of claim 161, wherein said structural or sequence information is obtained from MS/MS.

163. The method of claim 162, wherein a list of one or more biomolecules is provided to select an ion of a biomolecule for MS/MS analysis.

164. The method, of claim 163, wherein said list is an inclusion list.

165. The method of claim 163, wherein said list is an exclusion list.

166. The method of claim 161, wherein the structural or sequence information is used to identify a biomolecule from a database.

167. The method of claim 166, wherein the biomolecule is identified using a computer procedure selected from the group consisting of Mascot®, Protein Lynx Global Server, SEQUEST®/TurboSEQUEST, PepSEQ, SpectrumMill, or Sonar MS/MS.

168. The method of claim 166, wherein said database is the Genbank®, EMBL, NCBI, MSDB, SWISS-PROT®, TrEMBL, dbEST, or Human Genome Sequence database.

169. The method of claim 166, wherein in step (b) the ion is assigned to the biomolecule identified from said database.

170. The method of claim 143, wherein in step (b) the ion is assigned to the biomolecule by calculating an uncharged mass for the ion.

171. The method of claim 143, wherein in step (b) the ion is assigned to the biomolecule by peptide mass fingerprinting.

172. The method of claim 143, wherein in step (c) the integration is over one or more charge states, isotopes, scans, fragments of the biomolecule, fractions of a separation, or a combination thereof.

173. The method of claim 143, wherein said method further comprises calculating an abundance of the biomolecule relative to a control biological sample.

174. The method of claim 143, wherein said method further comprises calculating abundances of a plurality of biomolecules relative to a control biological sample.

175. The method of claim 173, wherein the abundance is used to diagnose a disease or condition.

176. The method of claim 173, wherein abundance is used to determine a biomolecule to target with a drug.

177. The method of claim 173, wherein an increase or decrease in abundance or the presence or absence of a biomolecule in the biological sample relative to the control biological sample is indicative of a disease or condition.

178. The method of claim 173, wherein the abundance is used to determine an amount of an isoform of a biomolecule.

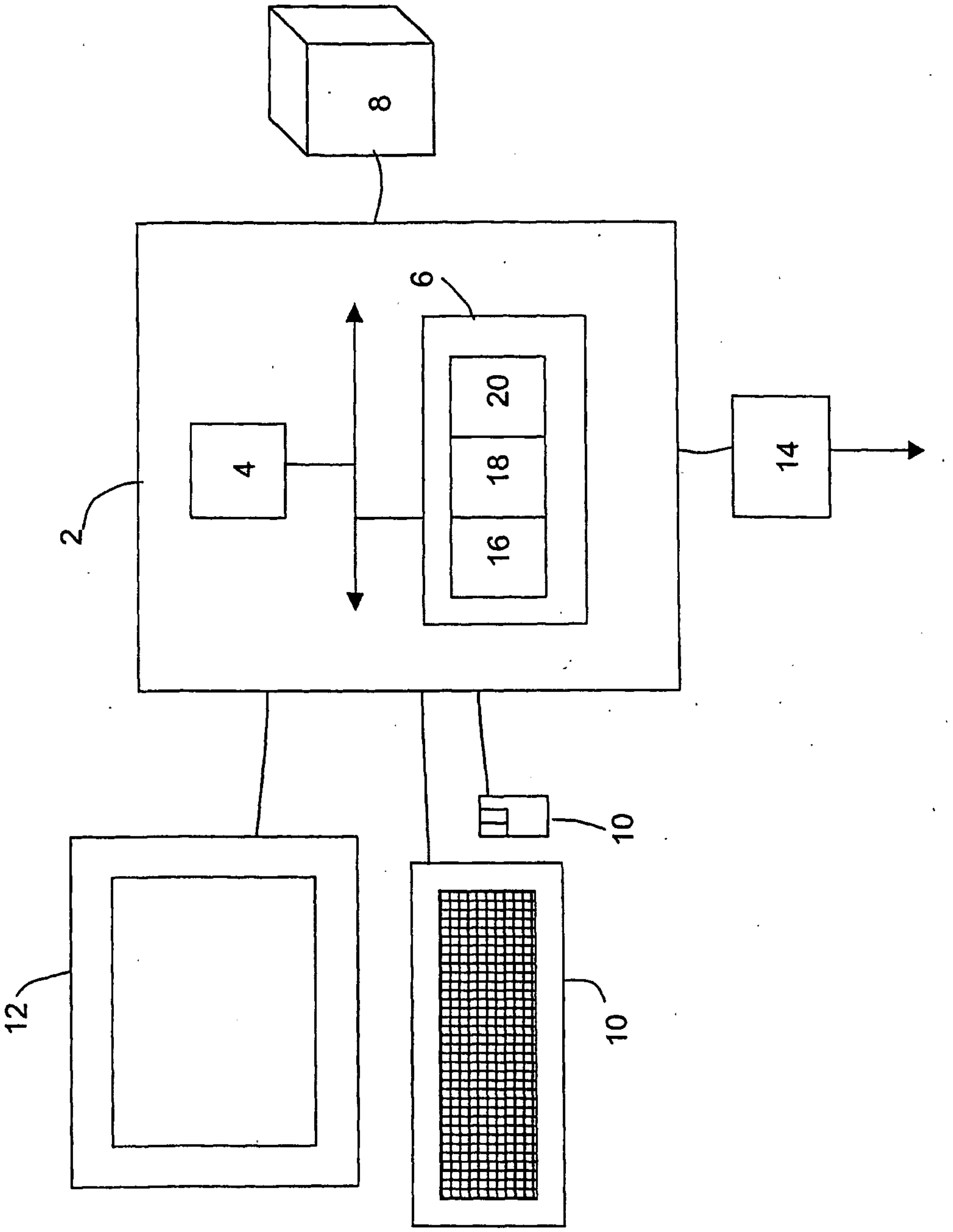


FIG. 1

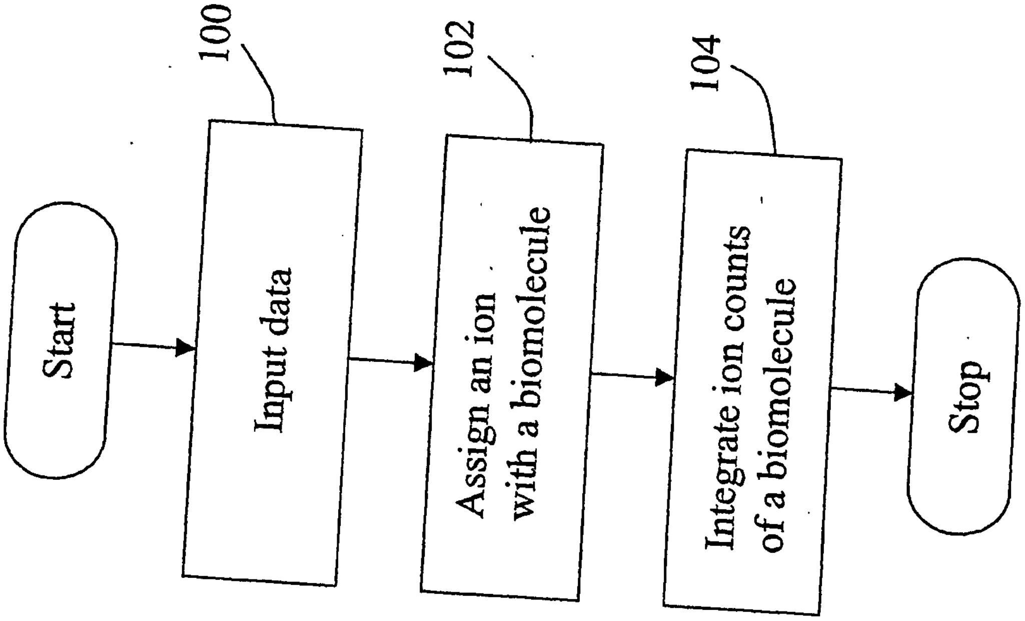

Start
Input data
100
Assign an ion with a biomolecule
102
Integrate ion counts of a biomolecule
104
Stop

FIG. 2

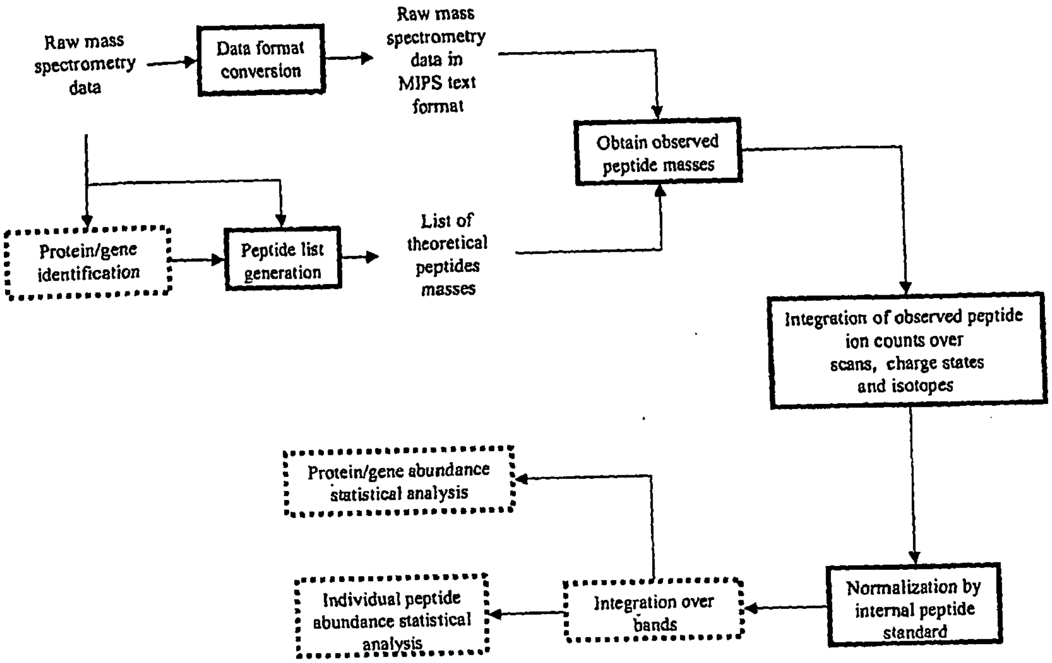
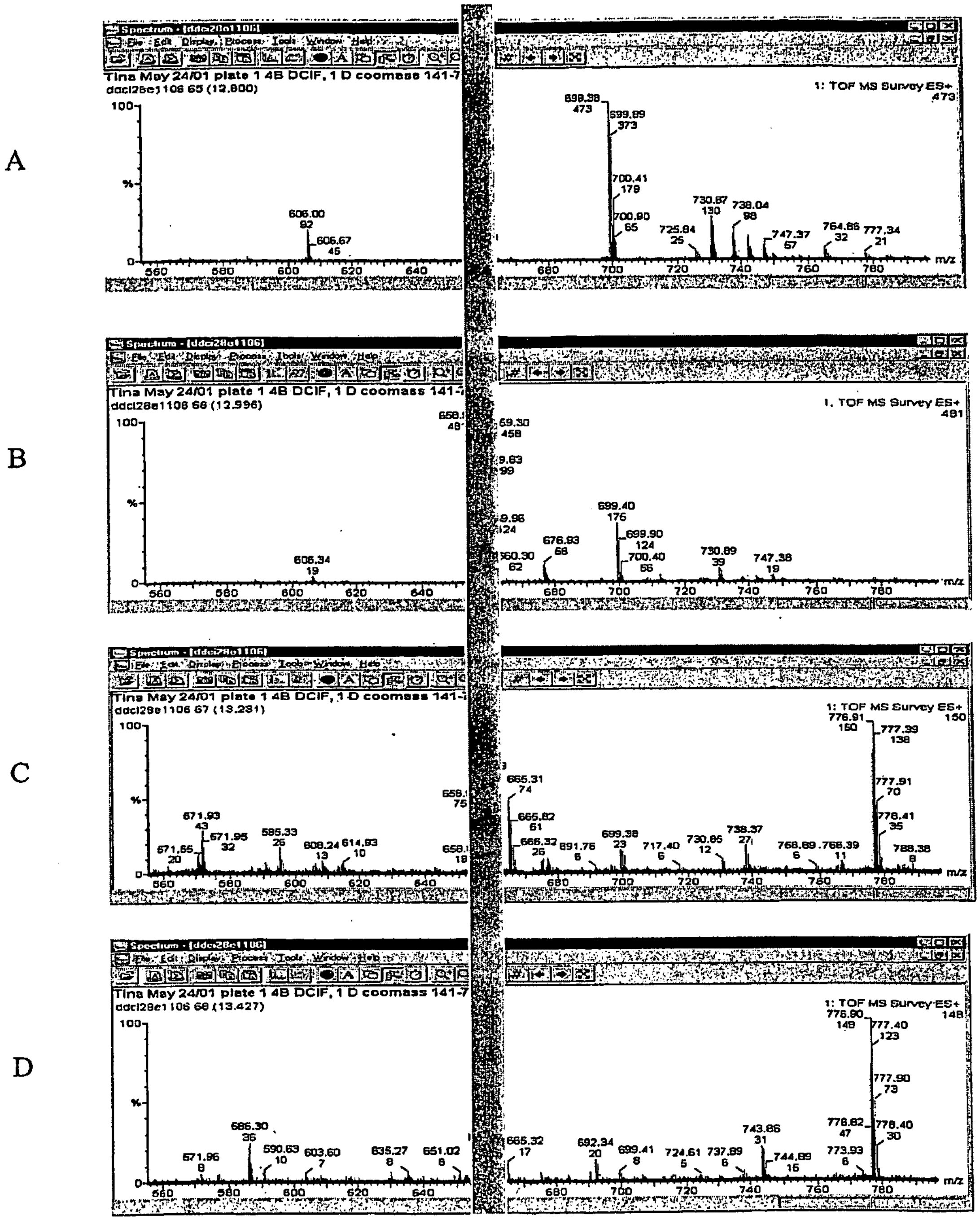
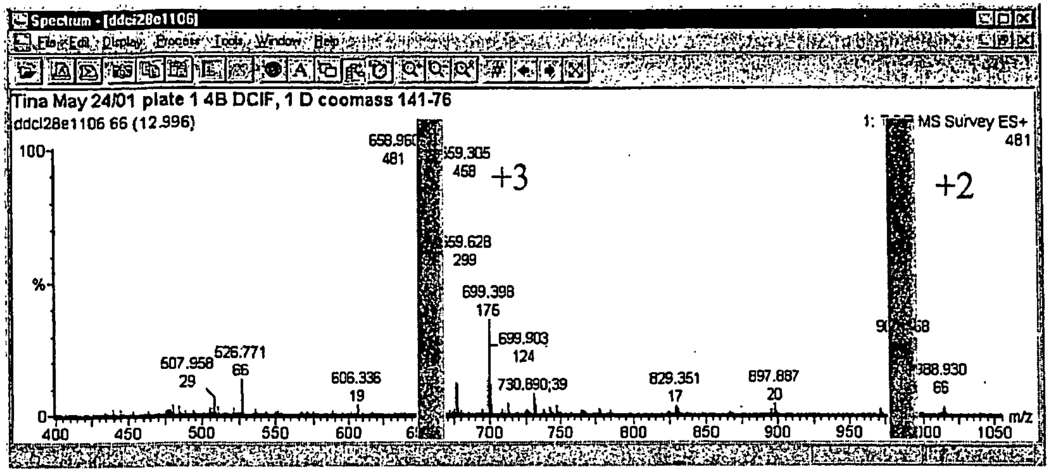

Raw mass spectrometry data
Data format conversion
Raw mass spectrometry data in MIPS text format
Obtain observed peptide masses
Protein/gene identification
Peptide list generation
List of theoretical peptides masses
Integration of observed peptide ion counts over scans, charge states and isotopes
Protein/gene abundance statistical analysis
Individual peptide abundance statistical analysis
Integration over bands
Normalization by internal peptide standard

FIG. 3

FIG. 4

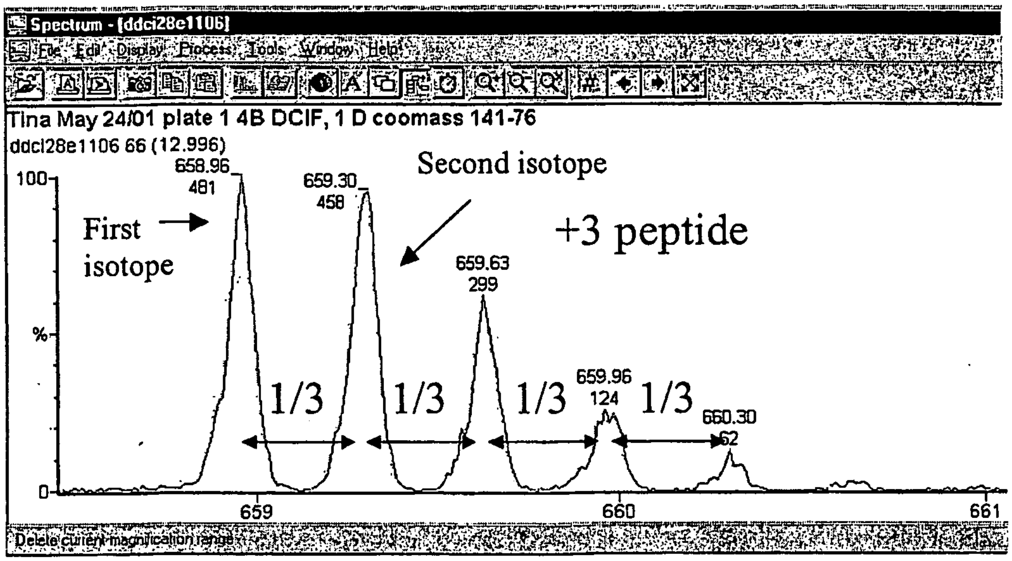

A
Spectrum - [ddcl28e1106]
File Edit Display Process Tools Window Help
Tina May 24/01 plate 1 4B DCIF, 1 D coomass 141-7
ddcl28e1106 65 (12.800)
1: TOF MS Survey ES+
473
100
%
0
698.38
473
699.89
373
700.41
179
700.90
65
725.84
25
730.87
130
738.04
98
747.37
57
764.86
32
777.34
21
606.00
82
606.67
45
560
580
600
620
640
680
700
720
740
760
780
m/z
B
Spectrum - [ddcl28e1106]
Tina May 24/01 plate 1 4B DCIF, 1 D coomass 141-
ddcl28e1106 66 (12.996)
1. TOF MS Survey ES+
481
100
%
0
658.
48
69.30
458
9.83
99
9.86
124
699.40
176
699.90
124
700.40
56
676.93
68
660.30
62
730.89
39
747.38
19
606.34
19
560
580
600
620
640
680
700
720
740
760
780
m/z
C
Spectrum - [ddcl28e1106]
Tina May 24/01 plate 1 4B DCIF, 1 D coomass 141-
ddcl28e1106 67 (13.231)
1: TOF MS Survey ES+
150
100
%
0
776.91
150
777.39
138
777.91
70
778.41
35
788.38
8
665.31
74
665.82
61
666.32
26
691.76
6
699.38
23
717.40
6
730.85
12
738.37
27
768.89
6
768.39
11
571.93
43
571.95
32
571.55
20
585.33
26
608.24
13
614.93
10
658.
75
658.
18
560
580
600
620
640
680
700
720
740
760
780
m/z
D
Spectrum - [ddcl28e1106]
Tina May 24/01 plate 1 4B DCIF, 1 D coomass 141-7
ddcl28e1106 68 (13.427)
1: TOF MS Survey ES+
148
100
%
0
776.90
148
777.40
123
777.90
73
778.82
47
778.40
30
773.93
6
743.86
31
744.89
15
737.89
6
724.61
5
699.41
8
692.34
20
665.32
17
571.96
8
586.30
36
590.63
10
603.60
7
635.27
8
651.02
8
560
580
600
620
640
680
700
720
740
760
780
m/z



Tina May 24/01 plate 1 4B DCIF, 1 D coomass 141-76

ddcl28e1106 66 (12.996)

FIG. 5



Tina May 24/01 plate 1 4B DCIF, 1 D coomass 141-76

ddcl28e1106 66 (12.996)

FIG. 6

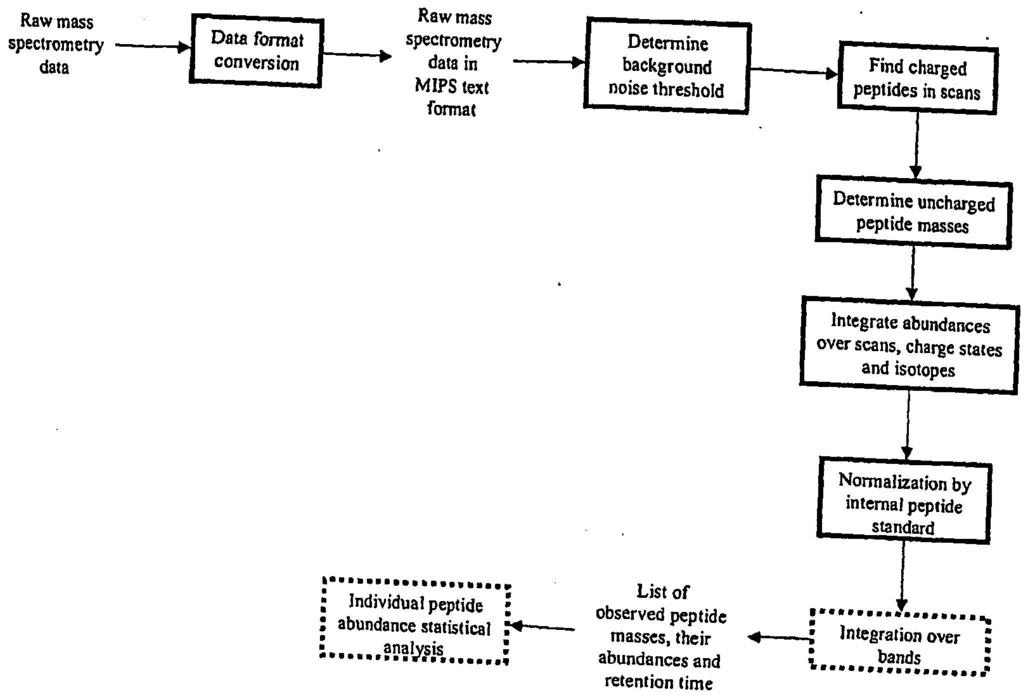
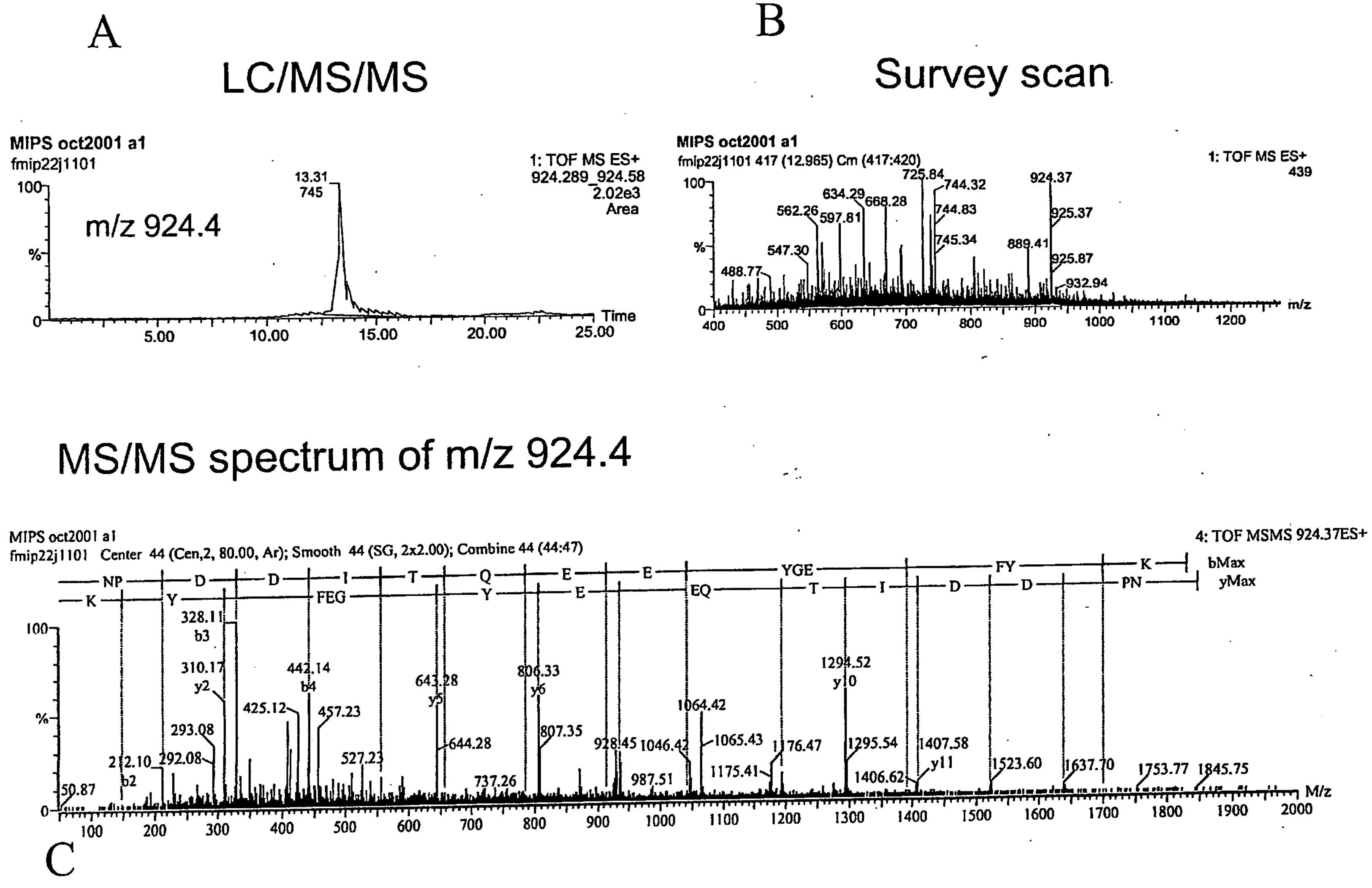

Raw mass spectrometry data
Data format conversion
Raw mass spectrometry data in MIPS text format
Determine background noise threshold
Find charged peptides in scans
Determine uncharged peptide masses
Integrate abundances over scans, charge states and isotopes
Normalization by internal peptide standard
Integration over bands
List of observed peptide masses, their abundances and retention time
Individual peptide abundance statistical analysis

FIG. 7

FIG. 8



A LC/MS/MS

B Survey scan

MS/MS spectrum of m/z 924.4

C

FIG. 8

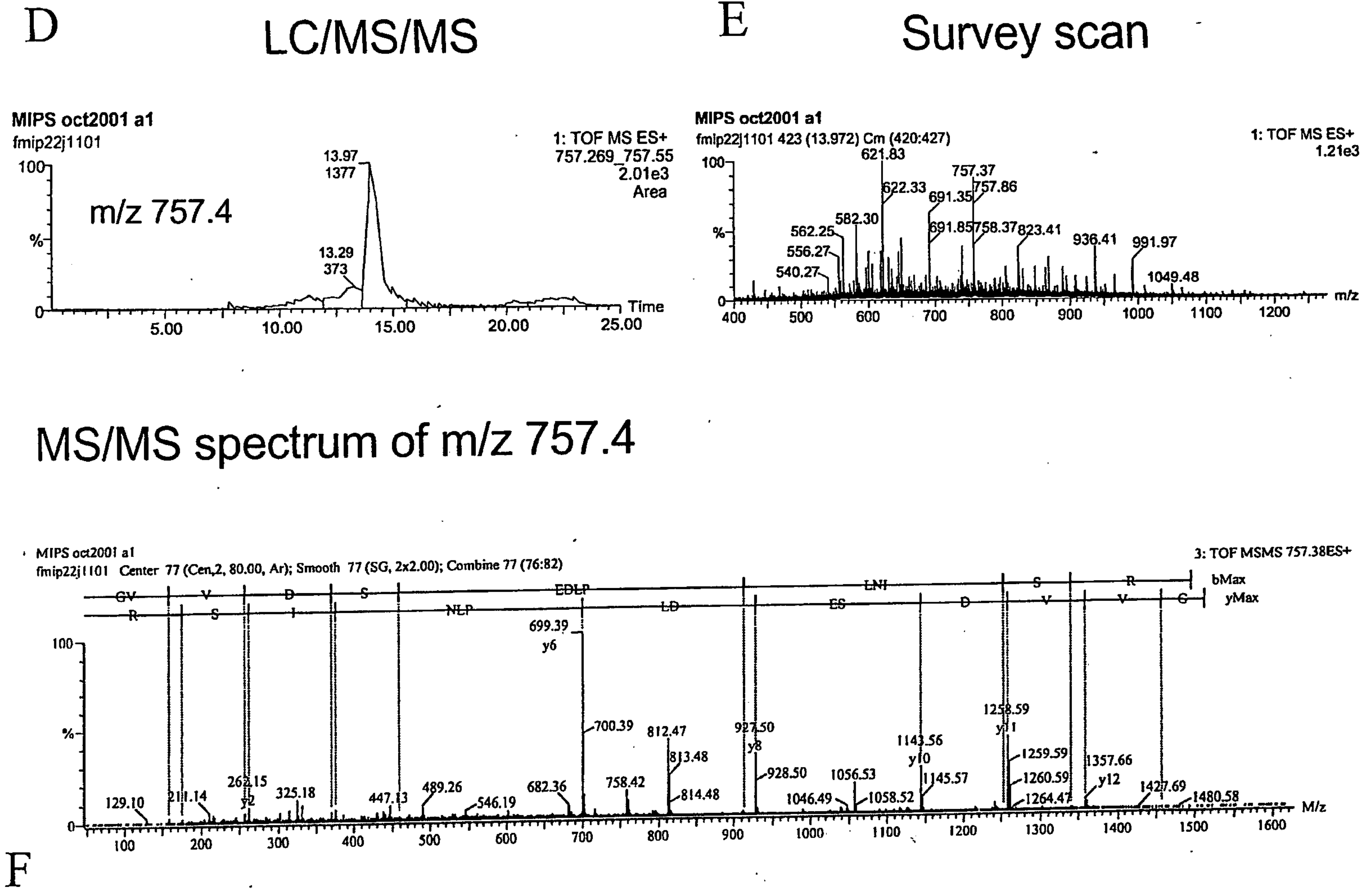

D
LC/MS/MS
MIPS oct2001 a1
fmip22j1101
1: TOF MS ES+
757.269_757.55
2.01e3
Area
m/z 757.4
13.97
1377
13.29
373
100
%
0
Time
5.00
10.00
15.00
20.00
25.00
E
Survey scan
MIPS oct2001 a1
fmip22j1101 423 (13.972) Cm (420:427)
1: TOF MS ES+
1.21e3
621.83
622.33
582.30
562.25
556.27
540.27
757.37
757.86
691.35
691.85
758.37
823.41
936.41
991.97
1049.48
100
%
0
m/z
400
500
600
700
800
900
1000
1100
1200
MS/MS spectrum of m/z 757.4
MIPS oct2001 a1
fmip22j1101 Center 77 (Cen,2, 80.00, Ar); Smooth 77 (SG, 2x2.00); Combine 77 (76:82)
3: TOF MSMS 757.38ES+
GV
V
D
S
EDLP
LNI
S
R
bMax
R
S
I
NLP
LD
ES
D
V
V
G
yMax
699.39
y6
700.39
812.47
813.48
814.48
927.50
y8
928.50
1143.56
y10
1145.57
1258.59
y11
1259.59
1260.59
1264.47
1357.66
y12
1427.69
1480.58
129.10
211.14
262.15
y2
325.18
447.13
489.26
546.19
682.36
758.42
1046.49
1056.53
1058.52
100
%
0
M/z
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
F

FIG. 8

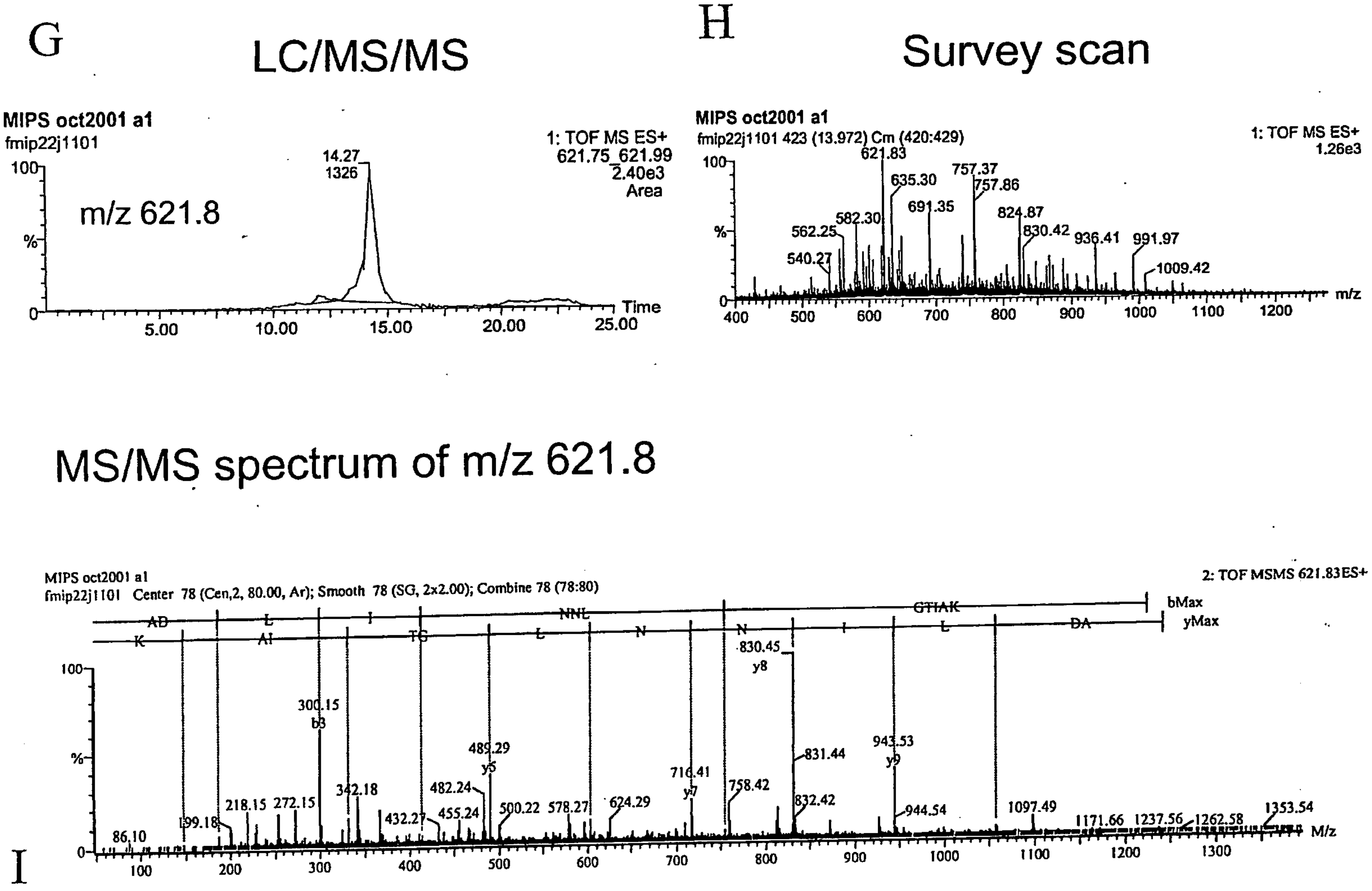
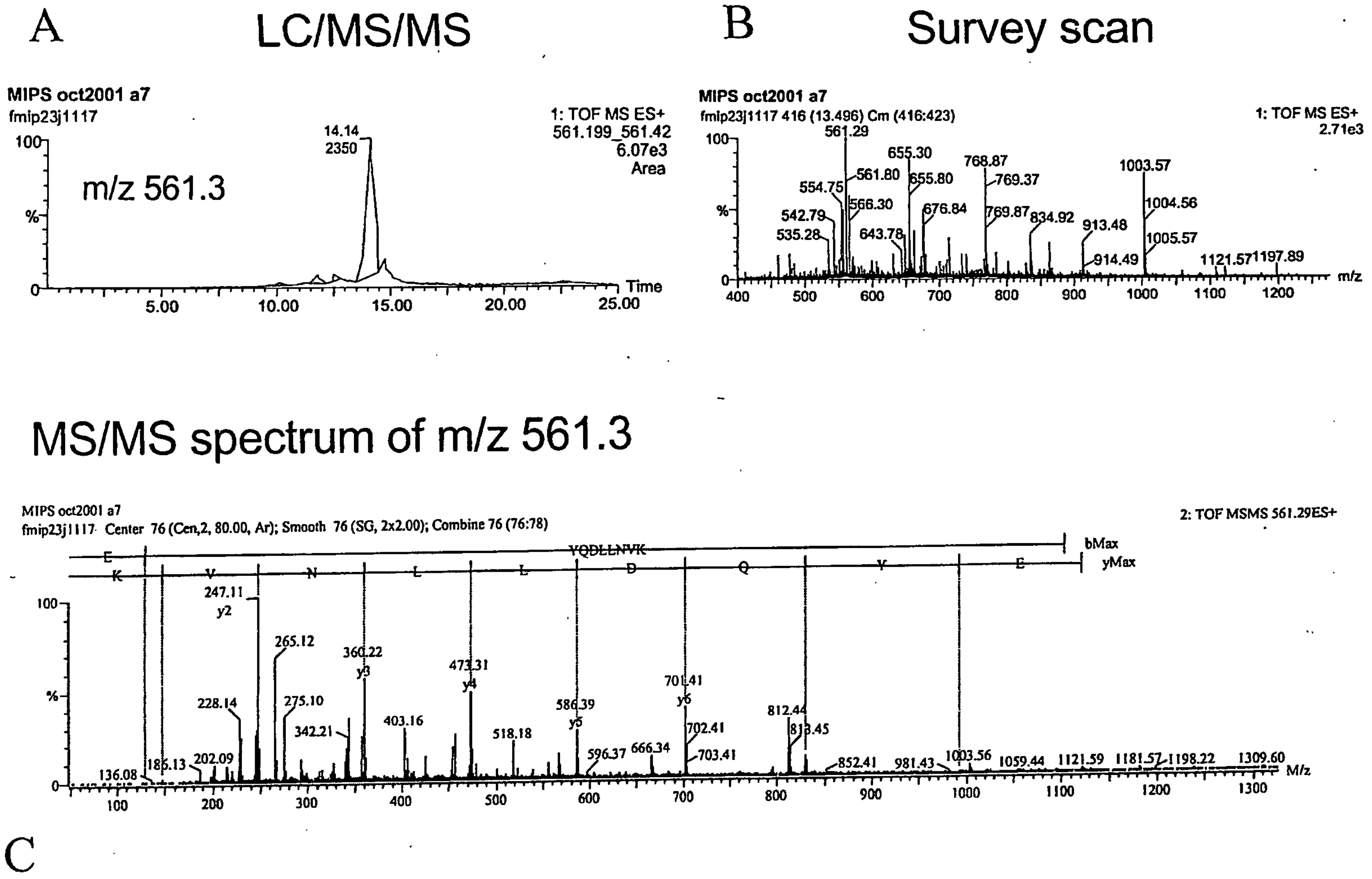

G
LC/MS/MS
MIPS oct2001 a1
fmip22j1101
1: TOF MS ES+
621.75_621.99
2.40e3
Area
m/z 621.8
14.27
1326
100
%
0
5.00
10.00
15.00
20.00
25.00
Time
H
Survey scan
MIPS oct2001 a1
fmip22j1101 423 (13.972) Cm (420:429)
1: TOF MS ES+
1.26e3
621.83
635.30
757.37
757.86
691.35
582.30
562.25
540.27
824.87
830.42
936.41
991.97
1009.42
100
%
0
400
500
600
700
800
900
1000
1100
1200
m/z
MS/MS spectrum of m/z 621.8
MIPS oct2001 a1
fmip22j1101 Center 78 (Cen,2, 80.00, Ar); Smooth 78 (SG, 2x2.00); Combine 78 (78:80)
2: TOF MSMS 621.83ES+
AD
L
I
NNL
GTIAK
bMax
K
AI
TG
L
N
N
I
L
DA
yMax
830.45
y8
300.15
b3
489.29
y5
716.41
y7
943.53
y9
831.44
86.10
199.18
218.15
272.15
342.18
432.27
455.24
482.24
500.22
578.27
624.29
758.42
832.42
944.54
1097.49
1171.66
1237.56
1262.58
1353.54
100
%
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
M/z
I

FIG. 9

Monocytes	Macrophages	Peak area		
(%)	(%)	m/z 924.4	m/z 757.4	m/z 621.8
100	0	77	355	234
75	25	44	217	198
50	50	27	197	104
25	75	15	74	34
0	100	0.4	5.3	0.0

FIG. 10



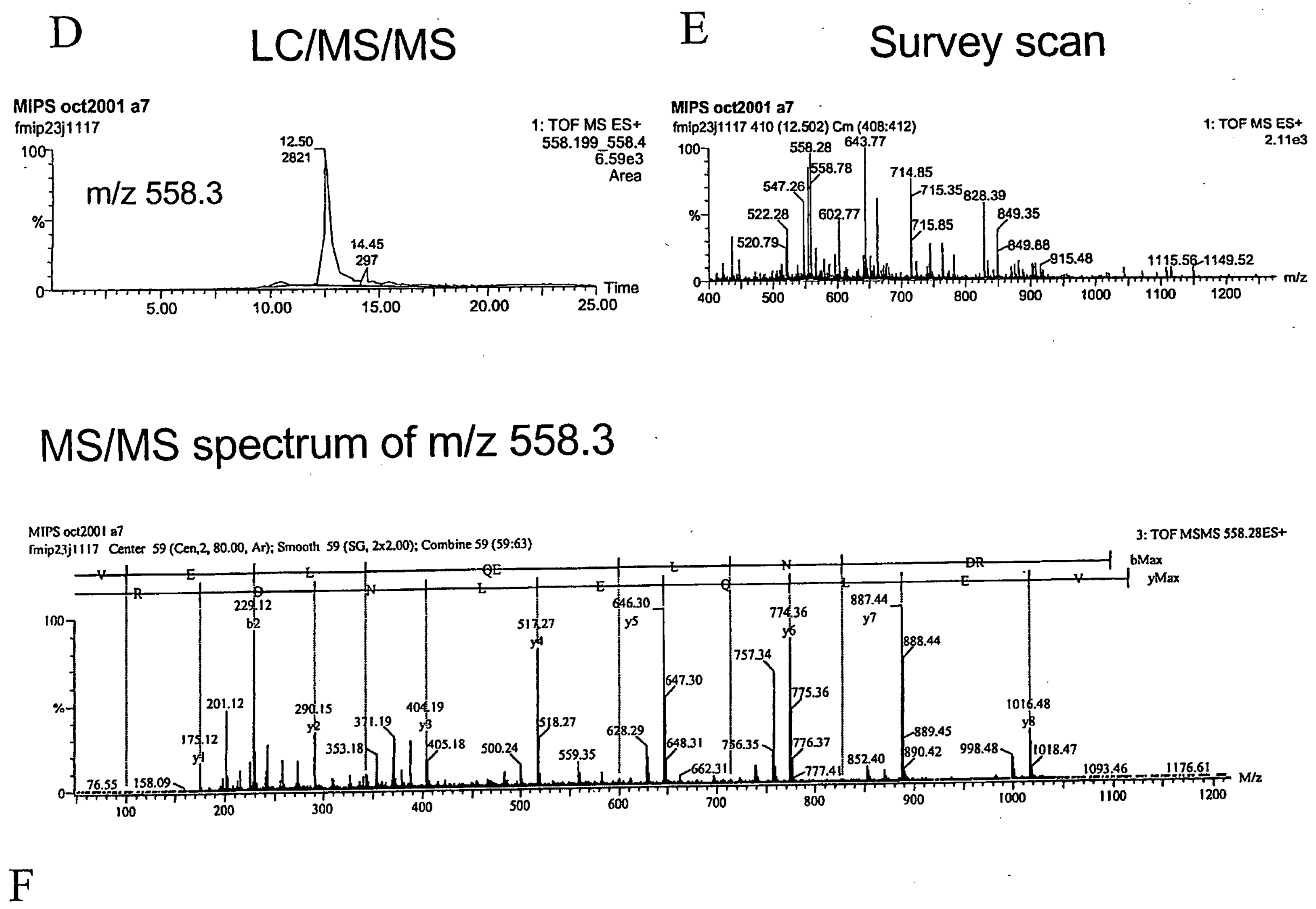
A LC/MS/MS

B Survey scan

MS/MS spectrum of m/z 561.3

C

FIG. 10



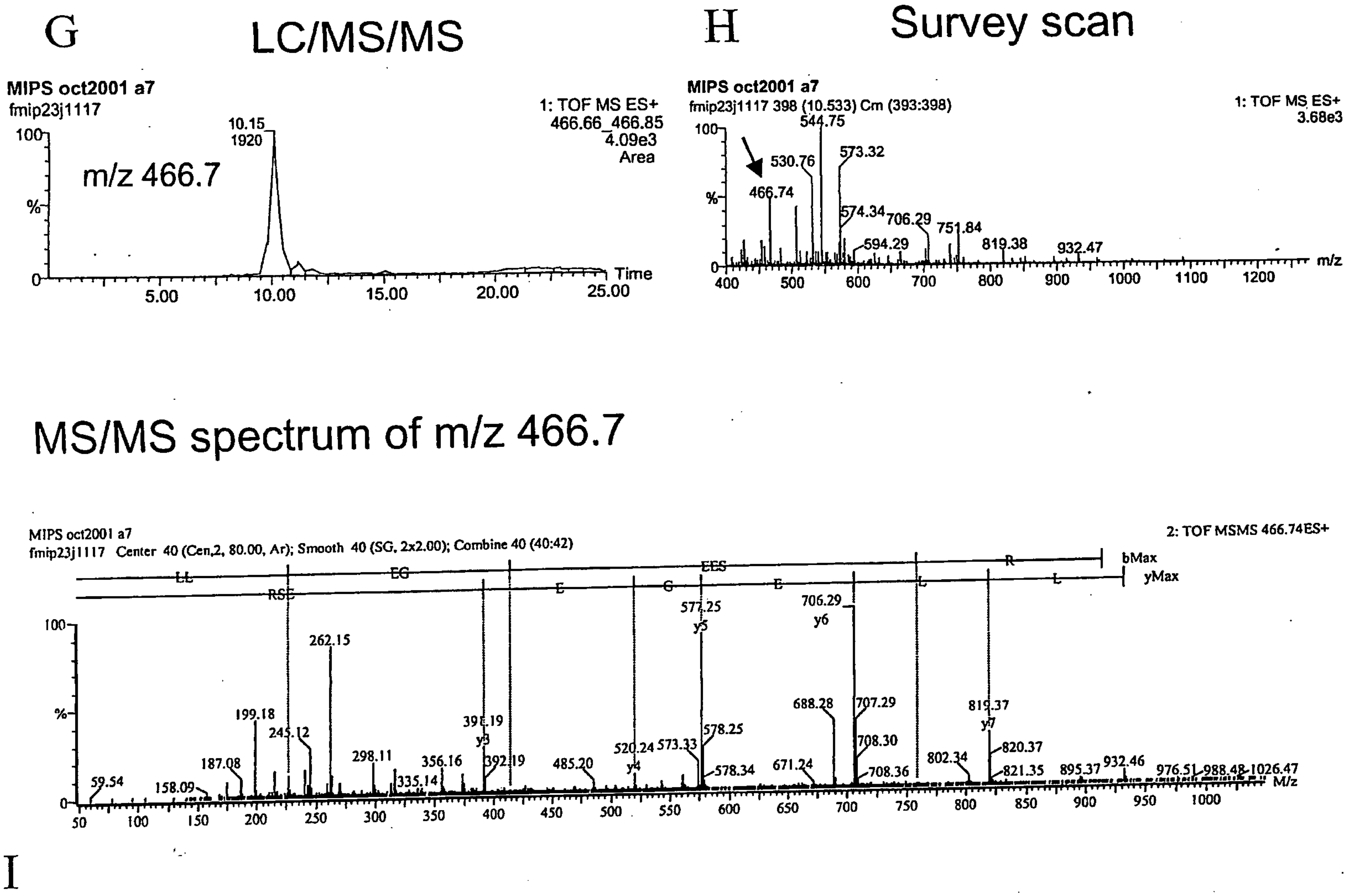
D LC/MS/MS

E Survey scan

MS/MS spectrum of m/z 558.3

F

FIG. 10



G LC/MS/MS

H Survey scan

MS/MS spectrum of m/z 466.7

I

FIG. 11

Monocytes	Macrophages	Peak area		
(%)	(%)	m/z 561.3	m/z 558.3	m/z 466.7
100	0	0	0	0
75	25	170	245	148
50	50	706	873	661
25	75	1098	1397	1044
0	100	1673	2309	1446

FIG. 12

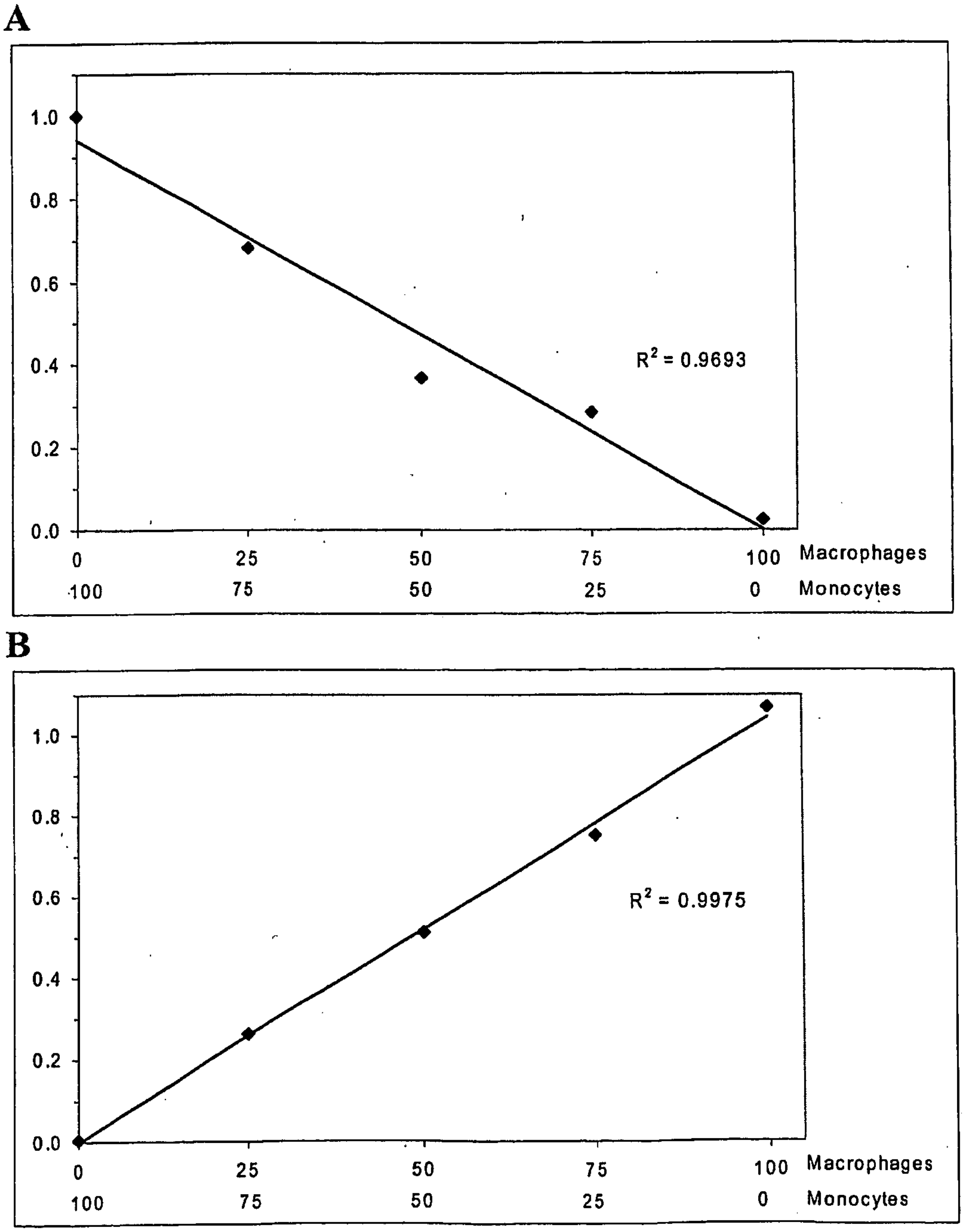

A
1.0
0.8
0.6
0.4
0.2
0.0
R² = 0.9693
0
25
50
75
100
Macrophages
100
75
50
25
0
Monocytes
B
1.0
0.8
0.6
0.4
0.2
0.0
R² = 0.9975
0
25
50
75
100
Macrophages
100
75
50
25
0
Monocytes

FIG. 13

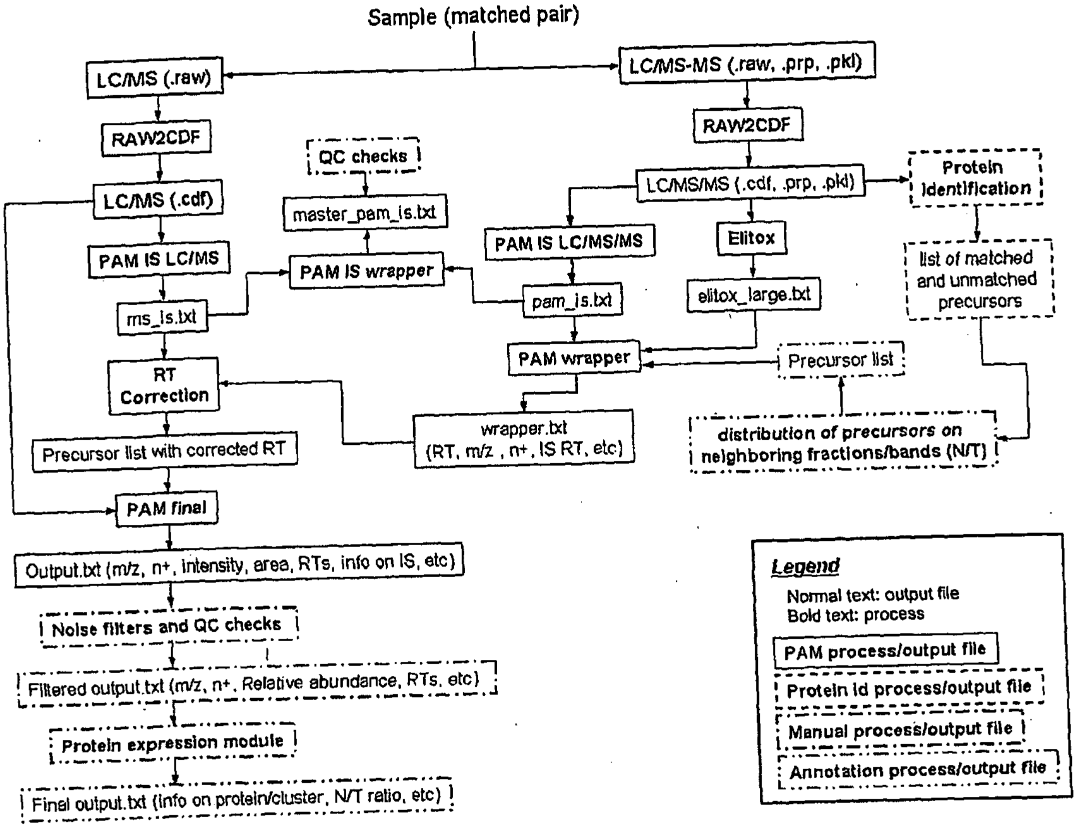
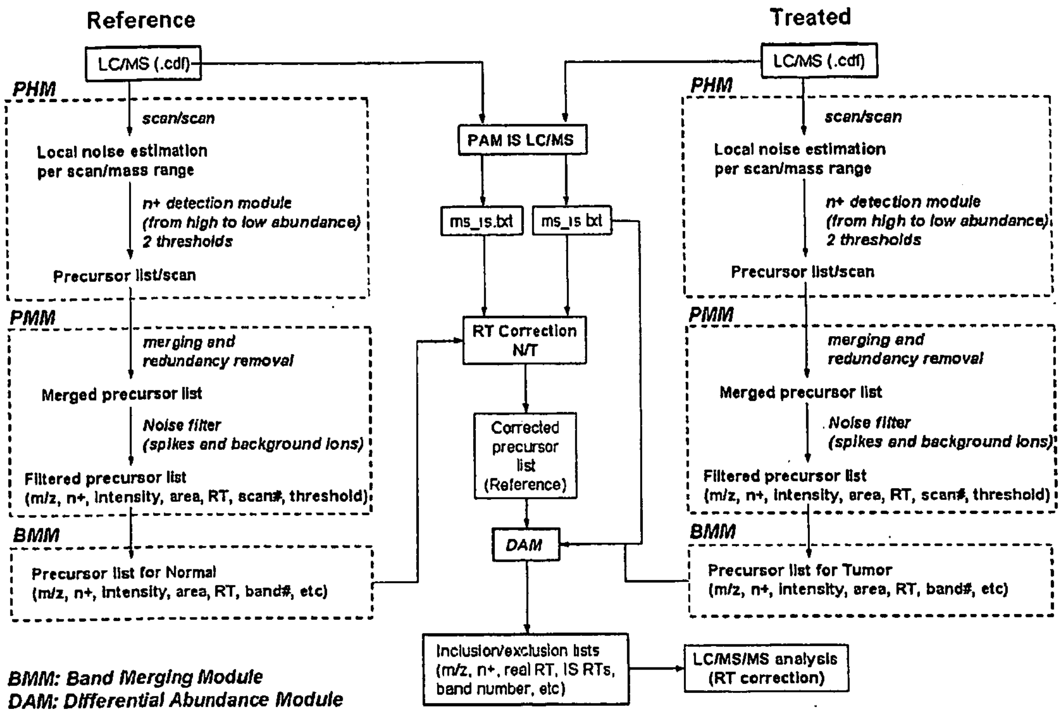

Sample (matched pair)
LC/MS (.raw)
LC/MS-MS (.raw, .prp, .pkl)
RAW2CDF
RAW2CDF
QC checks
LC/MS (.cdf)
master_pam_is.txt
LC/MS/MS (.cdf, .prp, .pkl)
Protein Identification
PAM IS LC/MS
PAM IS wrapper
PAM IS LC/MS/MS
Elitox
list of matched and unmatched precursors
ms_is.txt
pam_is.txt
elitox_large.txt
RT Correction
PAM wrapper
Precursor list
wrapper.txt (RT, m/z , n+, IS RT, etc)
distribution of precursors on neighboring fractions/bands (N/T)
Precursor list with corrected RT
PAM final
Output.txt (m/z, n+, intensity, area, RTs, info on IS, etc)
Noise filters and QC checks
Filtered output.txt (m/z, n+, Relative abundance, RTs, etc)
Protein expression module
Final output.txt (info on protein/cluster, N/T ratio, etc)
Legend
Normal text: output file
Bold text: process
PAM process/output file
Protein id process/output file
Manual process/output file
Annotation process/output file

FIG. 14

		Time allocation	Patient: 48 bands, 1 exclusion/band 50 000 precursors, Plattform of 24 CPUs
.raw			
↓	File conversion	9 min (LC-MS), 6 min LC-MS-MS (per CPU)	42 minutes
NetCDF			
↓	Load in memory	10 min/LC-MS, 7 min/LC-MS-MS (per CPU)	48 minutes
↓	Elitox	12s/MS-MS file (per CPU)	48 seconds
.elt (Rt, m/z, n+)			
↓	PAM (IS)	5 min/file (per CPU)	30 minutes
Rt_{IS1}, Rt_{IS2}			
↓	PAM wrapper	15 min (all bands)	15 min (all bands)
.raw			
↓	Rt Correction	5 min/target band (per CPU)	10 minutes
↓	PAM final	35 seconds/precursor (per CPU)	20.3 hours

Expected processing time: 23 hours

FIG. 15


Reference
Treated
LC/MS (.cdf)
LC/MS (.cdf)
PHM
scan/scan
Local noise estimation per scan/mass range
n+ detection module (from high to low abundance) 2 thresholds
Precursor list/scan
PMM
merging and redundancy removal
Merged precursor list
Noise filter (spikes and background ions)
Filtered precursor list (m/z, n+, intensity, area, RT, scan#, threshold)
BMM
Precursor list for Normal (m/z, n+, intensity, area, RT, band#, etc)
Precursor list for Tumor (m/z, n+, intensity, area, RT, band#, etc)
PAM IS LC/MS
ms_is.txt
ms_is.txt
RT Correction N/T
Corrected precursor list (Reference)
DAM
Inclusion/exclusion lists (m/z, n+, real RT, IS RTs, band number, etc)
LC/MS/MS analysis (RT correction)

BMM: Band Merging Module
DAM: Differential Abundance Module

FIG. 16

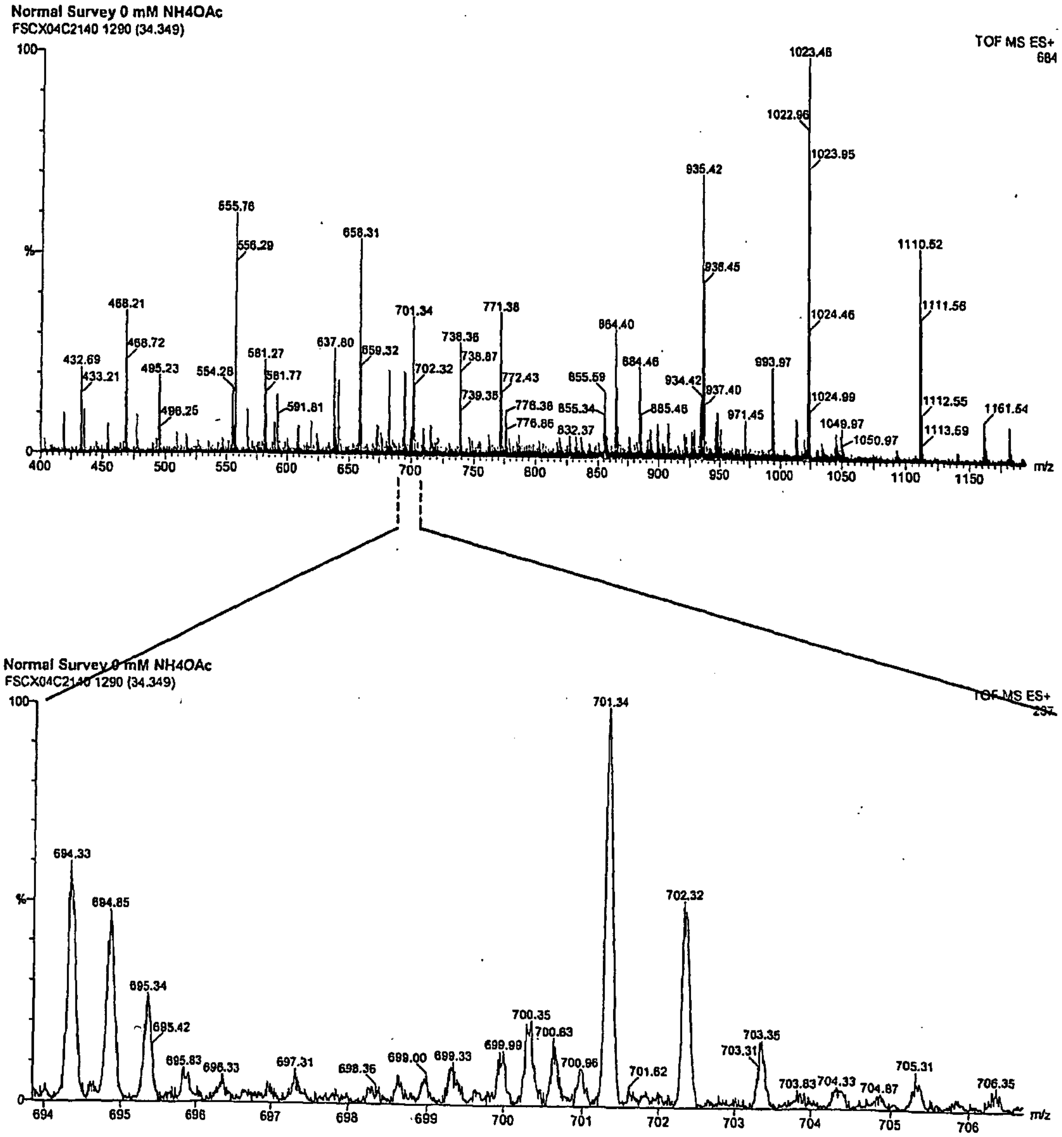

Normal Survey 0 mM NH4OAc
FSCX04C2140 1290 (34.349)
TOF MS ES+
684
100
%
0
1023.46
1022.96
1023.95
935.42
936.45
655.76
556.29
658.31
1110.52
1111.56
468.21
468.72
701.34
771.38
864.40
1024.46
432.69
433.21
495.23
554.28
581.27
581.77
637.80
659.32
738.36
738.87
702.32
772.43
855.59
884.46
934.42
937.40
993.97
1024.99
496.25
591.81
739.38
776.38
776.86
855.34
832.37
885.46
971.45
1049.97
1050.97
1112.55
1113.59
1161.64
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
1150
m/z
Normal Survey 0 mM NH4OAc
FSCX04C2140 1290 (34.349)
TOF MS ES+
297
100
%
0
701.34
694.33
694.85
702.32
695.34
695.42
695.83
696.33
697.31
698.36
699.00
699.33
699.99
700.35
700.63
700.96
701.62
703.35
703.31
703.83
704.33
704.87
705.31
706.35
694
695
696
697
698
699
700
701
702
703
704
705
706
m/z

FIG. 17

m/z	Charge	Abundance	m/z	Charge	Abundance	m/z	Charge	Abundance
418.6983	2	492	709.8170	3	450	875.3859	2	254
432.6915	2	1252	715.3726	1	660	876.4608	2	450
454.2169	2	444	718.3337	1	376	880.4158	1	321
468.2127	2	2349	720.3475	2	272	882.3663	2	292
477.2540	1	663	721.3445	1	337	884.4647	1	1801
493.2820	1	254	730.3380	1	280	890.4093	2	311
495.2327	1	1299	738.3555	2	2111	892.9196	2	552
509.2984	1	355	746.3610	1	353	898.8925	2	686
516.7577	2	257	749.4077	1	261	903.4128	1	326
546.7647	2	214	753.3340	2	247	907.4528	1	747
554.2811	1	1070	754.3984	1	321	915.4485	2	283
555.7617	2	3964	762.3549	2	376	920.4064	1	561
567.3177	1	936	765.4227	1	289	923.0950	3	256
581.2729	2	1648	771.3843	1	2599	926.4207	2	570
589.2795	1	638	775.8904	2	842	928.7324	3	453
591.3022	2	1080	779.3499	1	443	932.4172	1	296
600.2868	1	314	785.3995	1	315	933.9506	2	1251
609.3594	1	534	787.4251	1	475	935.4227	1	5116
619.3936	1	538	795.3463	1	383	938.4582	1	424
623.7786	2	316	798.3990	1	374	941.3857	2	155
637.7996	2	1776	803.3998	1	390	946.0868	3	630
641.3497	1	1482	806.4216	1	378	947.4304	2	838
658.3060	1	3739	817.3865	2	333	950.4222	2	537
661.3879	1	366	819.3595	1	474	970.9631	2	700
672.2948	2	347	825.3864	1	306	993.4849	2	1651
673.3321	1	554	827.3689	2	501	1013.5089	1	987
676.3641	1	604	832.3651	1	435	1019.9630	2	304
680.3499	1	341	836.4202	1	452	1022.9628	2	5500
682.3311	3	1347	843.4572	2	241	1034.4807	2	250
685.3512	1	343	848.4185	1	300	1045.5222	3	352
690.3279	1	340	850.4285	1	428	1049.4448	2	533
693.3978	2	426	855.5892	3	1036	1092.5658	1	238
694.3331	2	1571	857.4284	2	473	1110.5249	1	4108
699.9900	3	524	861.3883	2	226	1161.5369	1	897
701.3402	1	2630	864.3970	1	2604	1181.6085	1	706
705.3121	1	271	867.4473	1	409			

FIG. 18

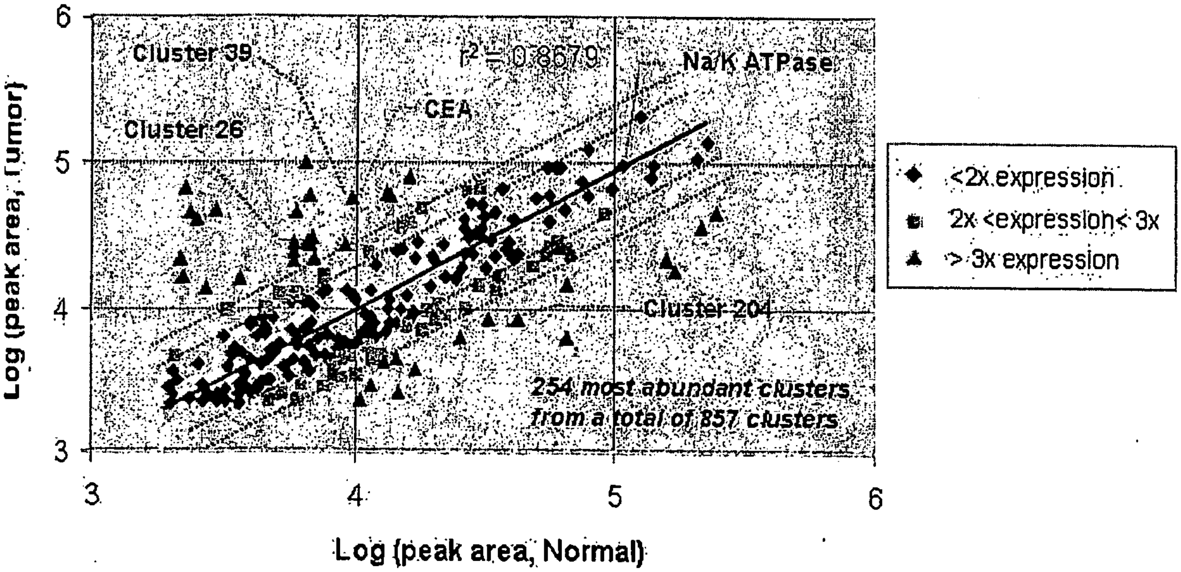
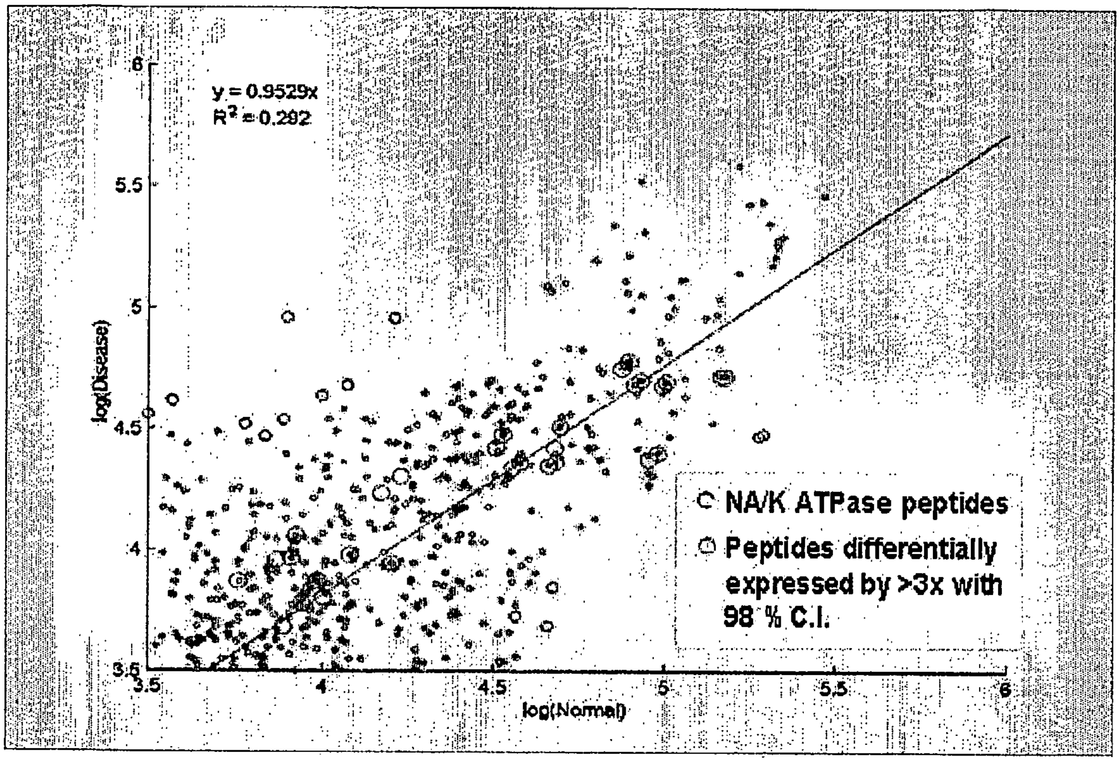

Cluster 39
$r^2 = 0.8679$
Na/K ATPase
CEA
Cluster 26
Cluster 204
254 most abundant clusters from a total of 857 clusters
◆ <2x expression
■ 2x <expression< 3x
▲ > 3x expression
Log (peak area, Tumor)
Log (peak area, Normal)
3
4
5
6
3
4
5
6

FIG. 19


y = 0.9529x
R^2 = 0.292
log(Disease)
log(Normal)
3.5
4
4.5
5
5.5
6
○ NA/K ATPase peptides
○ Peptides differentially expressed by >3x with 98 % C.I.